

Explainable AI in Medical Imaging and Diagnosis: A Critical Review of Methods, Clinical Translation Barriers, and an Integrated Evaluation Framework

Ali Alhazmi^{a, b, c} 

Abstract

Deep learning now matches or outperforms specialist clinicians on many medical imaging benchmarks, but clinical uptake has lagged well behind the benchmark numbers. Opacity is usually blamed. Explainable artificial intelligence (XAI) is the field's proposed fix, and it has produced a large and still-growing toolkit: saliency maps, feature-attribution scores, concept-based explainers, and counterfactual image generators, spanning radiology, digital pathology, dermatology, and ophthalmology. This critical review works through more than 35 recent primary studies, systematic reviews, and meta-analyses to test an assumption the literature rarely states outright but almost always relies on: that producing an explanation is the same thing as producing understanding. It is not, at least not reliably. Quantitative fidelity and stability data, human-factors evidence on automation bias and trust calibration, and the regulatory direction set by the US Food and Drug Administration and the EU Artificial Intelligence Act all point the same way: widely used *post-hoc* methods such as Grad-CAM, LIME, and SHAP routinely fail basic faithfulness checks, explanations can make clinicians more confident in wrong predictions rather than less, and the field currently evaluates all this along technical, psychological, and regulatory tracks that barely intersect. The review's central contribution is an integrated Clinical-XAI Evaluation Framework built around six dimensions, fidelity, robustness, plausibility, actionability, equity, and regulatory readiness, meant to give editors, developers, and clinical adopters one shared checklist for deciding whether an explanation has actually earned a place in a diagnostic workflow, rather than several separate and often contradictory ones. The review argues that the field should shift its default investment away

from *post-hoc* explainability and toward inherently interpretable architectures, modality-specific evaluation protocols, and prospective validation tied to decision outcomes, not further proliferation of attribution maps that look convincing but have rarely been checked against what the model actually did.

Keywords: Explainable artificial intelligence; Medical imaging; Clinical decision support; Diagnostic AI; Interpretability; Trust calibration; Radiology; Digital pathology

Introduction

Deep convolutional and transformer architectures now perform diagnostic tasks in radiology, pathology, dermatology, and ophthalmology at levels rivaling specialist clinicians, yet clinical adoption lags because these systems are opaque. Explainable artificial intelligence (XAI) is the field's proposed answer, but this review argues that the literature has matured technically far faster than critically: most studies report a plausible-looking heatmap without asking whether it is faithful to the network's actual reasoning. A 67-study meta-analysis found fidelity scores of only 0.38 for SHAP and 0.54 for Grad-CAM against the models they claimed to explain, versus 0.81 for LIME, though at higher cost and worse noise stability [1]. Following Rudin's distinction between explaining a black box after the fact and building an interpretable model from the start [2], this review organizes the XAI taxonomy, synthesizes modality-specific evidence, examines fidelity and human factors, covers the regulatory landscape, proposes a six-dimension Clinical-XAI Evaluation Framework, addresses cross-cutting barriers, and sets out research gaps before concluding. Sources came from structured searches of PubMed, IEEE Xplore, ACM Digital Library, ScienceDirect, and arXiv, prioritizing recent systematic reviews and quantitative reader studies, supplemented by foundational methods papers including Grad-CAM, LIME, SHAP, and Rudin's critique [2]. This is a critical review, not a systematic one, and makes no claim to exhaustive coverage.

Scope and source selection

Literature was identified through structured searches of Pub-

Manuscript submitted July 30, 2026, accepted August 13, 2026
Published online August 21, 2026

^aDepartment of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia

^bEngineering and Technology Research Center, Jazan University, PO Box 114, Jazan 82817, Saudi Arabia

^cCorresponding Author: Ali Alhazmi, Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia. Email: alihazmi@jazanu.edu.sa

doi: <https://doi.org/10.14740/aicm36>
AI in Clinical Medicine
2819-7437 (online)

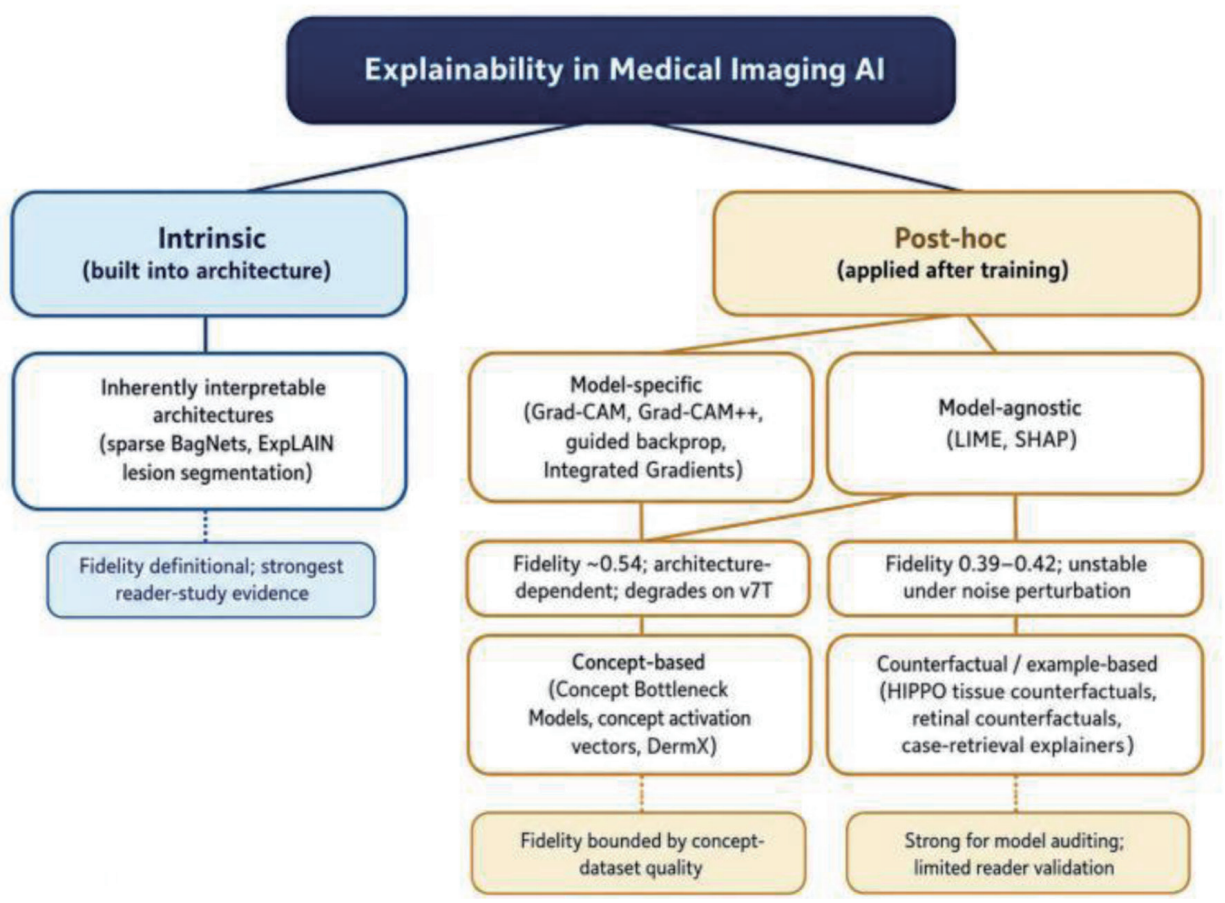


Figure 1. Taxonomy of explainable AI (XAI) methods applied to medical imaging, organized by the intrinsic/*post-hoc* and model-specific/model-agnostic axes described in Section “A Taxonomy of Explainability in Medical Imaging AI.” *Italicized notes summarize the fidelity evidence discussed in the text.*

Med, IEEE Xplore, ACM Digital Library, ScienceDirect, and arXiv (through July 2026), combining the terms “explainable AI,” “interpretability,” “saliency,” “Grad-CAM,” “LIME,” “SHAP,” “concept-based,” and “counterfactual” with “medical imaging,” “radiology,” “pathology,” “dermatology,” and “ophthalmology,” then extended by forward and backward citation tracing from the systematic reviews and meta-analyses these searches surfaced. Roughly 90 candidate sources were screened for relevance; about 40 met the inclusion priorities below and are cited. Three criteria governed inclusion: peer-reviewed systematic reviews and meta-analyses were prioritized over single primary studies where both existed on a topic; among primary studies, quantitative fidelity, stability, or reader-study outcomes were prioritized over purely qualitative demonstrations; and publication within the past 3 years was prioritized, with foundational methodological papers (Grad-CAM, LIME, SHAP, Rudin’s interpretability critique [2]) retained regardless of age as the direct ancestors of current practice. This is a critical review, not a PRISMA-guided systematic review, and inclusion was not governed by a pre-registered protocol; selection was guided by relevance to the review’s evaluative argument, and no claim is made that it captures every relevant study.

A Taxonomy of Explainability in Medical Imaging AI

Barredo Arrieta et al separate transparent models, whose logic a human can simulate directly, from opaque models needing a secondary explanatory mechanism, splitting the latter along intrinsic-versus-*post-hoc* and model-agnostic-versus-model-specific axes [3]. LIME [4] and SHAP [5] are model-agnostic; Grad-CAM [6] is model-specific, exploiting gradients into a convolutional network’s final feature maps. These distinctions carry real consequences: Grad-CAM run across five backbone architectures for lung cancer CT classification showed fidelity dropping sharply for Vision Transformers relative to convolutional networks [7], while LIME and SHAP, being sampling-based, can yield different explanations across repeated runs on the same image. Figure 1 lays out the full taxonomy. Saliency- and attribution-based methods dominate the field [8], descending from early gradient-based saliency maps [9], guided backpropagation [10], SmoothGrad [11], Integrated Gradients [12], and DeepLIFT [13], to Grad-CAM [6] and Grad-CAM++ [14]. Adebayo et al showed several popular saliency methods produce similar heatmaps even after model weights are

randomized, meaning they may not track whether the model learned anything at all [15], a finding the clinical XAI literature is only beginning to absorb as it shifts toward quantitative evaluation [16].

This section's later critique of *post-hoc* fidelity should not be read as a blanket dismissal, since in Section "The Fidelity Problem: Do Explanations Explain?", the picture's own meta-analytic evidence is uneven rather than uniformly negative: LIME's fidelity (0.81) substantially exceeded both SHAP's (0.38) and Grad-CAM's (0.54) in that same dataset [1], and Grad-CAM's failure mode is architecture-specific, holding reasonably well on convolutional networks while degrading mainly on transformers [7], not failing everywhere equally. *Post-hoc* methods also carry practical advantages the review's framing should not obscure: they attach to a model already trained and deployed, without the bespoke engineering, re-training, and architecture redesign that inherently interpretable models require; they are cheap enough to run at scale for routine developer-side auditing even where clinician-facing fidelity is weak; and SHAP's grounding in Shapley values gives it an axiomatic justification the purely visual saliency methods lack, even though that justification has not translated into the highest empirical fidelity scores here. The right reading of the evidence in this review is therefore that *post-hoc* method utility is genuinely application- and architecture-dependent, not that *post-hoc* explanation is categorically inferior.

Concept-based methods explain predictions via clinically meaningful concepts: Concept Bottleneck Models predict interpretable intermediate concepts before classification [17]; Lucieri et al used concept activation vectors for skin lesion explanation [18]; Patricio et al built concept-based explanations without *post-hoc* dependency [19] and later used vision-language models for the same purpose [20]; DermX trained concept localization end-to-end [21]. Their fidelity is bounded by how complete the typically small concept-annotated datasets are. Counterfactual explanation asks what minimal change would flip a prediction: HIPPO generates tissue-region counterfactuals in digital pathology for bias detection and model auditing [22]. Example-based explanation, retrieving similar cases, remains underused despite mapping naturally onto clinical reasoning [23]. Inherently interpretable architectures drop *post-hoc* explanation altogether: a sparse BagNet for diabetic retinopathy matched black-box accuracy (AUC 0.904) while its native evidence maps identified lesions with far higher precision (0.960) and improved screening speed [24]; ExplAIIn achieved comparable results via interpretable lesion segmentation [25]. This challenges the assumed accuracy-interpretability trade-off, put at a 5–7% AUC penalty on average in meta-analytic data [1] but not borne out by these two cases.

These architectures nonetheless carry costs the strength of the diabetic retinopathy evidence should not obscure. Both the sparse BagNet and ExplAIIn were purpose-built and extensively engineered for a single, well-defined task with mature, large public datasets [24, 25]; neither result demonstrates that an equivalent architecture can be built as easily, or would perform as well, for a task without that same depth of prior data and engineering investment, and the review is not aware of comparable head-to-head evidence outside ophthalmology. Building

an interpretable-by-design model typically requires committing to the architecture, and to what counts as an interpretable unit (a patch, a lesion segment, a concept), before training even begins, which is a heavier upfront design commitment than attaching a *post-hoc* explainer to a network already optimized for accuracy alone, and it forecloses the option of reusing an existing, validated black-box model. Tooling, pretrained backbones, and developer familiarity for intrinsically interpretable architectures also remain far less mature than for standard convolutional or transformer classifiers, a practical adoption barrier independent of any accuracy consideration. The honest summary is that inherently interpretable models have earned their strongest evidence in a narrow set of well-resourced use cases, and generalizing that evidence to modalities and tasks without comparable investment remains, at present, an assumption rather than a demonstrated result.

Modality-Specific Critical Synthesis

Radiology has generated the largest volume of medical imaging XAI research, yet adoption of explainability analysis remains limited, with saliency-based visualization dominating what little is done [26, 27]. A widely discussed case of an FDA-cleared algorithm misdiagnosing intracranial hemorrhage underscored that human-machine interaction, not clearance alone, determines safety, feeding calls for a human-AI symbiosis model of practice [28]. Digital pathology faces its own problem: gigapixel whole-slide images are tiled into patches, breaking biological continuity; reviewers argue explainability must be paired with causability, genuine causal understanding, not mere statistical association [29]. HIPPO's tissue-region counterfactuals offer a methodologically mature response, applied successfully to breast metastasis detection and glioma mutation classification [22]. Dermatology benefits from the pre-existing ABCDE vocabulary; Chanda et al's 116-dermatologist study found an explainable AI system reaching 81% balanced accuracy while significantly raising diagnostic confidence and trust [30], though the evidence still leans on non-clinical imagery with limited skin-tone diversity, and classifier performance, from Esteva et al's original network [31] through dermoscopy ensembles [32], has outrun explanation validation. Ophthalmology offers the clearest case for intrinsic interpretability: *post-hoc* methods show the worst noise-stability degradation of any modality (53% for SHAP versus 11% in radiology [1]), while inherently interpretable architectures match black-box accuracy with no *post-hoc* stage at all [24, 25]. Table 1 summarizes these four modalities.

The Fidelity Problem: Do Explanations Explain?

Fidelity, whether an explanation accurately reflects a model's true decision process, is not the same as plausibility, whether it looks reasonable to a human. A saliency map can highlight a visible lesion, and be entirely unfaithful to the network's actual computation. Figure 2 sketches this dissociation conceptually.

The 67-study meta-analysis found LIME's fidelity (0.81)

Table 1. Modality-Specific Summary of Dominant XAI Approach, Strongest Available Evidence, and Key Limitation

Modality	Dominant XAI approach	Strongest evidence identified	Key limitation/gap
Radiology	Saliency-based (Grad-CAM); adoption of any explainability analysis remains limited overall	Professional society (RSNA) advocacy for stronger transparency and post-deployment monitoring requirements.	A substantial share of published radiology AI studies report no explainability analysis at all [26, 27].
Digital pathology	Counterfactual tissue-region perturbation (HIPPO)	HIPPO validated across breast metastasis detection, melanoma/ breast cancer prognostication, and glioma mutation classification [22].	Patch-based tiling breaks biological continuity; explainability must be paired with causability, not treated as equivalent to it [29].
Dermatology	Concept-based methods; text- and region-based explanations	116-dermatologist, three-phase reader study: explainable AI significantly raised diagnostic confidence and trust versus an unexplained system [30].	Evidence base still leans on non-clinical dermoscopic imagery, with limited attention to skin-tone representation.
Ophthalmology	Inherently interpretable architectures (sparse BagNet, ExplAIIn)	Matched black-box accuracy with no post-hoc stage; highest-precision native evidence maps of any modality reviewed [24, 25].	<i>Post-hoc</i> methods show the worst noise-stability degradation of any modality (53% for SHAP) [1].

meaningfully ahead of SHAP’s (0.38) and Grad-CAM’s (0.54), directly contradicting common assumptions about gradient-based explanations that nonetheless keep Grad-CAM the field’s default [1]. The same work coined the term “explain-

ability trap” for post-hoc explanations creating an illusion of understanding that can measurably increase clinician overconfidence in wrong predictions. Figure 3 plots these three scores with their 95% confidence intervals.

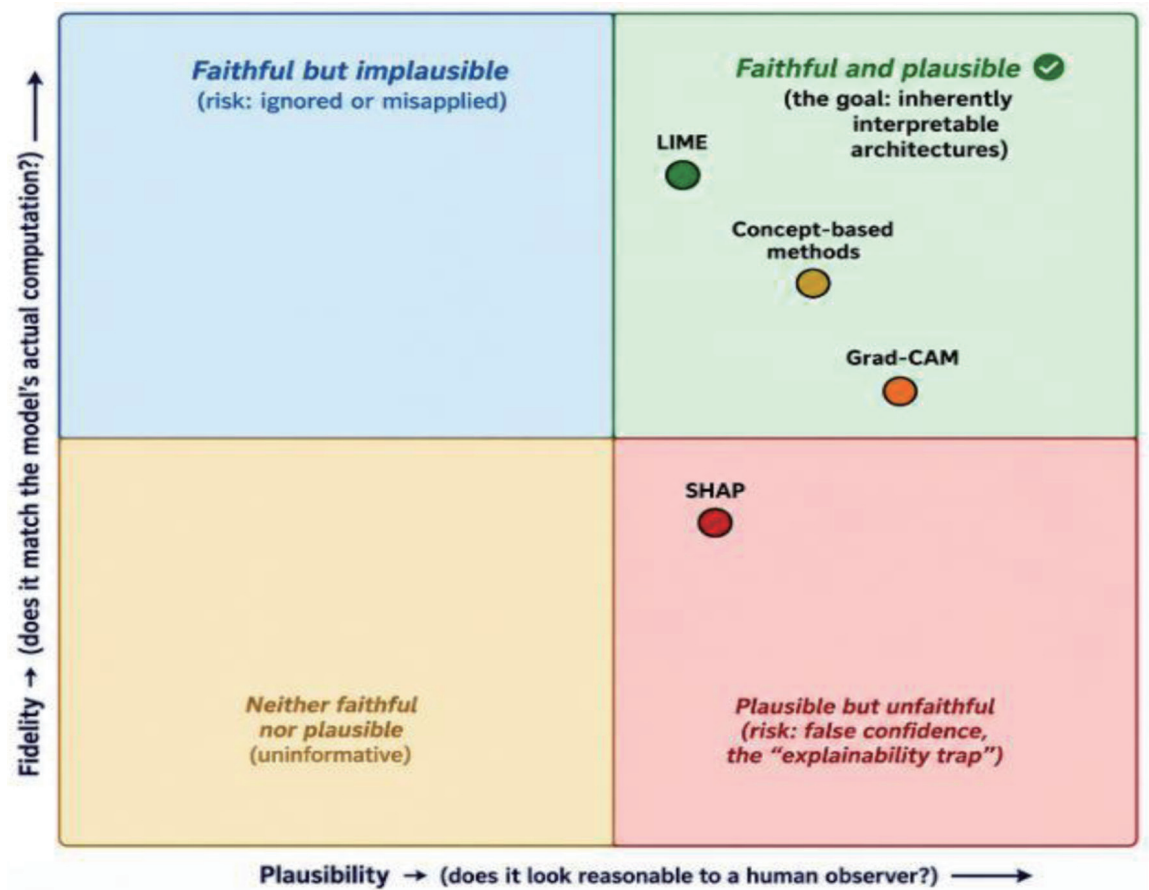


Figure 2. Fidelity versus plausibility: a conceptual illustration of the review’s central argument. Positions are qualitative, based on the evidence discussed in the text, not derived from a single quantitative scale.

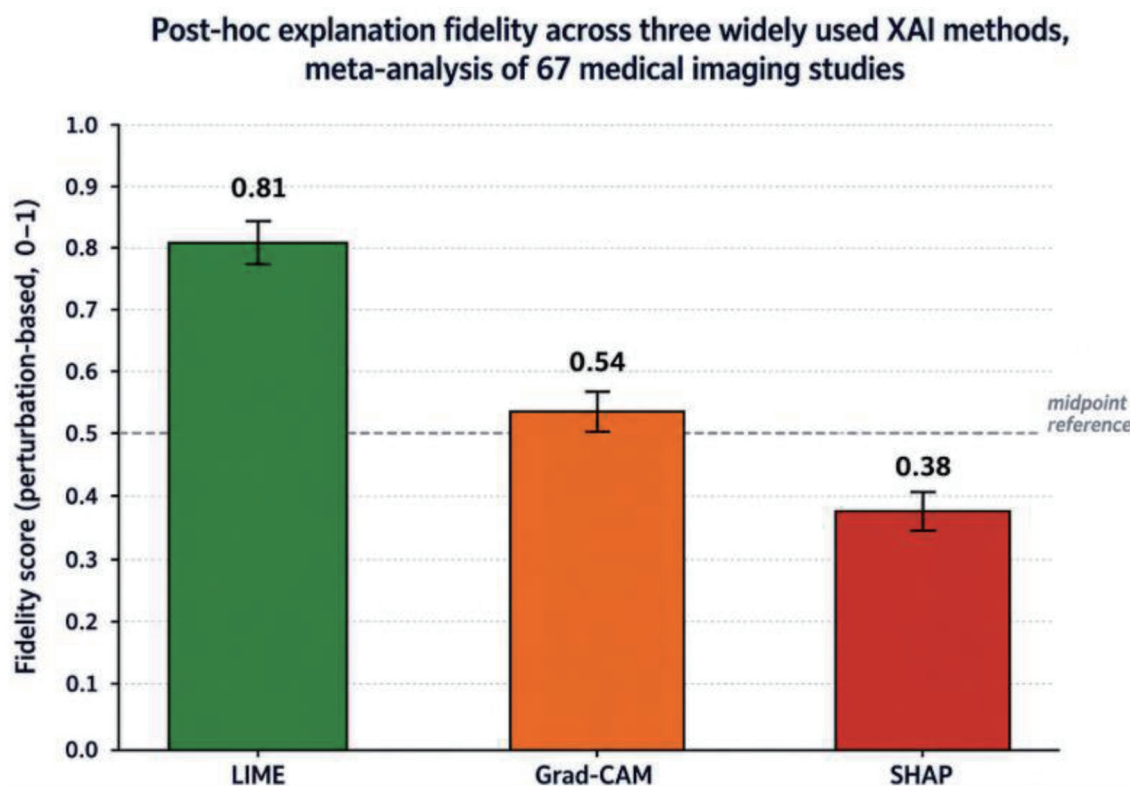


Figure 3. Perturbation-based fidelity scores (mean, 95% CI) for LIME, Grad-CAM, and SHAP, pooled across 67 studies spanning radiology, pathology, and ophthalmology applications. Source: meta-analytic evidence discussed in Section “The Fidelity Problem: Do Explanations Explain?” [1].

Architecture dependence compounds the problem: Grad-CAM’s fidelity dropped sharply for Vision Transformers relative to convolutional networks in controlled comparison [7], meaning a method validated on one architecture cannot be assumed faithful on another. Adebayo et al’s sanity-check test, randomizing model weights or labels and checking whether the explanation changes accordingly, remains an underused but essential minimum bar any clinical saliency method should clear before publication [15], consistent with the field’s broader, still-incomplete shift toward quantitative evaluation [16].

Human Factors: Trust, Automation Bias, and the Limits of Explanation

Automation bias, deferring uncritically to a wrong recommendation, and algorithm aversion, rejecting correct advice after one error, both recur in this literature; work on physician reliance in cancer management found evidence of both in the same population depending on how evidence was framed. Explanations do not reliably fix either and can worsen both: overly technical or overly simplistic explanations reinforce misplaced trust, especially among less experienced clinicians, while contrastive and hybrid explanations reduce automation bias more effectively than simple feature-based ones. Trust calibration research frames the goal as matching trust to actual reliability rather than maximizing it [33–35]; a 41-practitioner study

testing four explanation categories across correct and incorrect scenarios confirmed this matters in practice [36], and miscalibrated AI confidence display was found to increase distrust and disuse regardless of direction. Chanda et al’s dermatology study remains the clearest positive counter-example: aligned explanation measurably raised trust and confidence [30], suggesting the benefit is conditional on quality and clinical alignment, not automatic, and largely extrapolated to imaging from other domains by analogy rather than direct demonstration.

Regulatory and Governance Landscape

The EU Artificial Intelligence Act classifies diagnostic imaging AI as “high-risk,” requiring conformity assessment, human-in-the-loop oversight, and post-market monitoring with dataset traceability, effectively mandating explainability. In the US, the FDA’s review process for AI/ML-based Software as a Medical Device has had to build new mechanisms for algorithms updated after clearance; professional societies such as RSNA have urged stronger transparency guidance. Guidelines for evaluating clinical explainable AI argue existing techniques often fail to meet the documentation and validation standards regulators expect, pointing to the absence of a unified assessment protocol as the central barrier to harmonization [37], the finding that most directly motivates the next section. Most of this evidence comes from North American

Table 2. Regulatory Posture Toward Explainability in Diagnostic Imaging AI Across Three Jurisdictions

Jurisdiction	Governing framework	Risk classification for diagnostic imaging AI	Explainability requirement
European Union	EU Artificial Intelligence Act	“High-risk”	Conformity assessment before market approval; mandated human-in-the-loop oversight; post-market monitoring with dataset traceability; restrictions on fully automated decisions without human oversight.
United States	FDA review process for AI/ML-based Software as a Medical Device	Case-by-case clearance	No single explicit explainability mandate; professional societies (e.g., RSNA) have formally urged stronger transparency, explainability, and lifecycle-management guidance.
Gulf Cooperation Council/MENA	National digital-health and AI strategies (varies by country)	Not yet formally codified for diagnostic imaging AI specifically	Not yet codified; regulatory infrastructure is, in most cases, still developing relative to clinical AI investment.

and European systems; Gulf Cooperation Council and wider MENA health systems are investing heavily in AI-assisted imaging while regulatory infrastructure still catches up, so fidelity, robustness, and equity all need re-establishing locally rather than assumed from imported models. Table 2 compares the three regulatory postures.

Toward an Integrated Clinical-XAI Evaluation Framework

The literature evaluates explanations along whichever single axis a given research team specializes in, technical fidelity, human trust, or regulatory compliance, with no protocol requiring all three before a method is considered clinically ready. This review proposes an integrated Clinical-XAI Evaluation Framework across six dimensions, summarized in Figure 4.

Dimension 1, Fidelity, requires sanity checks against weight or label randomization [15] and perturbation-based faithfulness metrics [1] before a method is described as an “explanation” without qualification. Dimension 2, Robustness and Stability, requires consistency under noise and repeated runs, re-established per modality given documented variation (53% versus 11% SHAP degradation across modalities) [1]. Dimension 3, Plausibility, operationalizes the reader-study evidence reviewed in Section “Human Factors: Trust, Automation Bias, and the Limits of Explanation”: it requires blinded expert judgment of clinical alignment, following the model Chanda et al demonstrated in dermatology [30], and is assessed separately from fidelity since a method can score well on one and poorly on the other. Dimension 4, Actionability, requires evidence that the explanation changes clinical behavior and outcomes, not just stated confidence or the trust-calibration measures discussed in Section “Human Factors: Trust, Automation Bias, and the Limits of Explanation”, per the diabetic retinopathy screening-speed evidence [24]. Dimension 5, Equity, requires consistent performance across patient subgroups, given the shortcut-learning risks discussed in Section “Cross-Cutting Barriers”. Dimension 6, Regulatory Readiness, directly operationalizes Section “Regulatory and Governance Landscape”: it requires evidence the

method would satisfy the EU AI Act’s high-risk conformity, human-oversight, and post-market monitoring requirements and evolving FDA Software as a Medical Device guidance, including under model updates. These six work as a structured checklist rather than a single composite score, since a method may reasonably be strong on some dimensions and weak on others depending on intended use. Table 3 gives each dimension an operational definition and an example measurable criterion, addressing the concern that a six-dimension label without a stated metric is not, on its own, actionable for a reviewer or procurement committee.

Applying the framework: a worked example

Applied to the sparse BagNet diabetic retinopathy model [24]: Fidelity is satisfied by construction; Robustness is plausible but untested against systematic noise; Plausibility is strong, matching ophthalmologist-identified lesions; Actionability is well satisfied via direct screening-speed measurement; Equity evidence is thin; Regulatory Readiness is uncertain under continuous monitoring. Applied to a generic Grad-CAM chest radiograph classifier: Fidelity fails by default absent a study-specific check, given the 0.54 meta-analytic base rate [1]; Robustness is doubtful given documented transformer degradation [7]; Plausibility is often high, which is precisely what makes an unchecked fidelity gap dangerous; Actionability and Equity are rarely measured; Regulatory Readiness is uncertain for lack of fidelity documentation. Table 4 lays the comparison out dimension by dimension.

Comparative overview of major XAI method families

Table 5 draws together the method families discussed across Sections “A Taxonomy of Explainability in Medical Imaging AI” through “The Fidelity Problem: Do Explanations Explain?” into a single comparison, summarizing representative techniques, the fidelity evidence behind each, and the primary clinical risk each family carries.

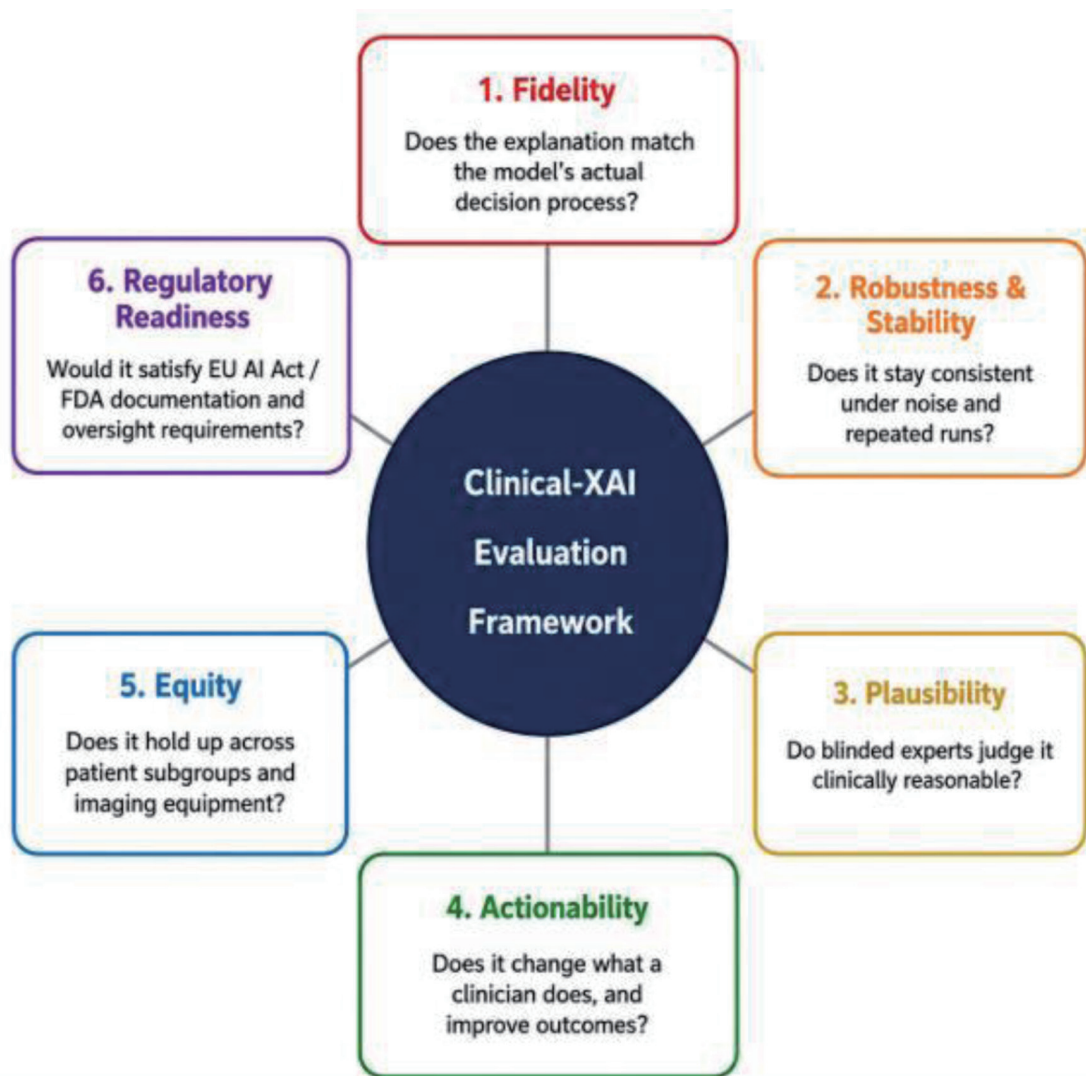


Figure 4. The six-dimension Clinical-XAI Evaluation Framework proposed in Section "Toward an Integrated Clinical-XAI Evaluation Framework". Each dimension is assessed independently rather than combined into a single composite score; see Section "Applying the framework: a worked example" for a worked example applying all six to two contrasting cases.

Cross-Cutting Barriers

Shortcut learning, networks exploiting spurious correlations (scanner signature, demographic proxies) rather than genuine pathology, cuts across every modality discussed. Work across 13 datasets spanning X-rays, CTs, ECGs, clinical text, and auscultation data found performance overestimated by up to 20% due to hidden acquisition biases invisible to standard validation. A faithful explanation of such a model will correctly flag the shortcut as important, but a clinician may misread it as genuine pathology, so explainability at best exposes the shortcut and at worst lends it unearned credibility. Federated learning introduces its own tension: a review of privacy-preserving federated learning found sensitive information can still be inferred from exchanged gradients, and cross-site heterogeneity creates fairness challenges of its own [38]. Explainability-federated learning integration

remains underdeveloped, with early work sketching what it requires [39] and federated explainable AI for Alzheimer's prediction one of few direct attempts [40]. Finally, workflow burden is real: perturbation-based methods need many forward passes per prediction, a cost that multiplies under robustness testing, and any added latency outside existing PACS workflows risks being skipped regardless of merit, favoring architectures whose explanation is a free byproduct of prediction itself.

Research Gaps and Future Directions

Five priorities stand out. Quantitative fidelity and stability benchmarking remains the exception rather than the norm [1, 7] and should be a required component of methods papers. Outcome-based actionability evidence, of the kind the diabetic retin-

Table 3. Operational Definitions and Example Measurable Evaluation Criteria for the six Clinical-XAI Evaluation Framework Dimensions

Dimension	Operational definition	Example measurable criterion	Evidence this dimension draws on
1. Fidelity	Degree to which the explanation reflects the model's actual computation, not merely a plausible-looking approximation of it	Passes a weight/label-randomization sanity check [15]; perturbation-based faithfulness score reported with a stated threshold (e.g., ≥ 0.6) [1]	Section "The Fidelity Problem: Do Explanations Explain?"
2. Robustness & stability	Consistency of the explanation under small, clinically irrelevant input perturbations and repeated stochastic runs	$\geq 20\%$ change in explained region under $\geq 10\%$ input noise, tested per modality rather than assumed from another task [1]	Section "The Fidelity Problem: Do Explanations Explain?"
3. Plausibility	Alignment of the explanation with clinically recognized reasoning, judged independently of fidelity	Blinded expert-rater agreement (e.g., Cohen's $\kappa \geq 0.6$) that the highlighted evidence matches recognized diagnostic criteria [30]	Section "Human Factors: Trust, Automation Bias, and the Limits of Explanation"
4. Actionability	Whether the explanation measurably changes clinician behavior or downstream outcomes, not only stated confidence	Pre-registered outcome (e.g., reading time, diagnostic accuracy, or referral appropriateness) shown to improve under AI support with the explanation present [24]	Section "Human Factors: Trust, Automation Bias, and the Limits of Explanation"
5. Equity	Consistency of both model performance and explanation quality across patient subgroups and acquisition equipment	Subgroup performance gap (e.g., across skin tone, sex, age, scanner type) reported and below a pre-specified threshold, not merely aggregate accuracy	Section "Cross-Cutting Barriers"
6. Regulatory readiness	Whether the explanation methodology and its evidence would satisfy applicable jurisdictional documentation and oversight requirements	Explicit mapping of the evidence above to EU AI Act Article-level conformity requirements and/or FDA SaMD guidance, including a stated plan for re-validation after model updates	Section "Regulatory and Governance Landscape"

opathy study provided [24], is rare outside ophthalmology and dermatology. Most of the evidence itself, reader studies, fidelity benchmarks, regulatory analyses, comes from North American, European, and East Asian settings and should not be assumed to transfer unvalidated elsewhere. Architecture transfer is barely studied beyond the single comparison in Section "The Fidelity Problem: Do Explanations Explain?", a growing problem as transformers displace convolutional backbones. Equity-stratified evaluation of explanation quality across skin tone, sex, age, and imaging equipment is largely absent despite documented shortcut-learning risks. And the intersection of federated learn-

ing, explainability, and continuous model updates under evolving post-market monitoring obligations remains genuinely open.

Conclusion

The medical imaging XAI literature has built an enormous technical toolkit but has done far less to establish which tools, for which task, under which validation standard, actually deliver trustworthy AI-assisted diagnosis. The finding that widely deployed methods like SHAP and Grad-CAM achieve fidelity

Table 4. The Six-Dimension Clinical-XAI Evaluation Framework Applied to Two Contrasting Cases, as Worked Through in Section "Applying the Framework: A Worked Example"

Dimension	Sparse BagNet for diabetic retinopathy screening [24]	Generic Grad-CAM for chest radiograph triage
1. Fidelity	Satisfied by construction—the evidence map is the model's actual computation, not an approximation	Fails by default without a study-specific fidelity check; meta-analytic base rate is only 0.54 [1]
2. Robustness & stability	Plausible given the architecture, but never systematically stress-tested against noise	Doubtful—fidelity documented to degrade sharply on transformer backbones [7]
3. Plausibility	Strong—evidence regions matched ophthalmologist-identified lesions with high precision	Tends to be high, but this is exactly what makes an unchecked fidelity gap dangerous
4. Actionability	Well satisfied—screening speed and accuracy under AI support measured directly [24]	Rarely measured in studies that report only qualitative heatmaps
5. Equity	Thin—static, single-institution training	Rarely assessed in typical Grad-CAM studies
6. Regulatory readiness	Uncertain—unclear how evidence maps hold up under continuous post-market monitoring	Uncertain—no fidelity documentation exists to build a regulatory case on

Table 5. Comparative Overview of Major XAI Method Families Discussed in This Review

Method family	Representative techniques	Fidelity evidence	Primary clinical risk
Gradient/saliency-based (post-hoc, model-specific)	Grad-CAM, Grad-CAM++, guided backpropagation, Integrated Gradients	Moderate and architecture-dependent; fidelity ~0.54 for Grad-CAM, degrades further on transformer backbones [1, 7]	Plausible-looking but unfaithful heatmaps may foster false confidence
Perturbation-based (post-hoc, model-agnostic)	LIME, SHAP	Variable: LIME ~0.81, SHAP ~0.38 in meta-analytic evidence; both unstable under noise, especially SHAP in ophthalmology [1]	Explanation instability across repeated runs undermines medico-legal defensibility
Concept-based	Concept Bottleneck Models, concept activation vectors, DermX	High plausibility where concept annotations are clinically grounded; fidelity bounded by concept-dataset quality [17, 18, 21]	Concept vocabulary incompleteness or annotation bias limits generalizability
Counterfactual/example-based	HIPPO tissue counterfactuals, retinal image counterfactual generation, case-retrieval explainers	Strong for model auditing and bias detection; limited large-scale clinical reader validation to date [22, 23]	Synthetic counterfactual images may not represent clinically realistic disease presentations
Inherently interpretable architectures	Sparse BagNet, ExplAIIn lesion-segmentation model	Explanation is definitionally faithful to computation; strongest available reader-study evidence of clinical benefit [24, 25]	Architecture constraints may limit applicability to tasks beyond well-engineered use cases

well below their ubiquity would suggest, and that explanations can increase clinician overconfidence in wrong predictions, should temper the assumption that adding an explanation module to a black box constitutes progress. The stronger evidence, from diabetic retinopathy screening and validated concept-based dermatology systems, favors building interpretability into architecture from the outset, reserving *post-hoc* explanation for where intrinsic interpretability cannot yet match required performance, subject to the structured evaluation proposed in Section “Toward an Integrated Clinical-XAI Evaluation Framework”. As EU and US regulatory frameworks move toward mandating transparency and human oversight, the field now has both empirical grounds and regulatory incentive to treat interpretability as a design requirement, not an afterthought.

Acknowledgments

None to declare.

Financial Disclosure

This review received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares no conflict of interest.

Author Contributions

As sole author, Ali Alhazmi conceived the review, conducted

the literature synthesis, developed the Clinical-XAI Evaluation Framework, drafted the manuscript, and approved the final version.

Data Availability

The author declares that data supporting the findings of this study are available within the article. As a narrative critical review, no new primary data were generated; the “data” underlying this study’s conclusions are the cited literature, all of which is publicly available through the publishers or preprint servers listed in the reference list.

AI Use Declaration

During preparation of this manuscript, the author used Claude (Anthropic) to assist with literature search, synthesis of source material, and drafting of the manuscript text. The author directed the research question, evaluation framework, and critical argument; reviewed, edited, and take full responsibility for all content, citations, interpretations, and conclusions presented. All cited sources were independently verified against their original publications. No AI tool is listed as an author.

Abbreviations

AI: artificial intelligence; XAI: explainable artificial intelligence; ML: machine learning; CAM: Class Activation Mapping (as in Grad-CAM); LIME: Local Interpretable Model-agnostic Explanations; SHAP: SHapley Additive exPlanations; AUC: area under the (receiver operating characteristic) curve; CI: confidence interval; CT: computed tomography; MRI:

magnetic resonance imaging; ECG: electrocardiogram; EU: European Union; FDA: US Food and Drug Administration; SaMD: Software as a Medical Device; RSNA: Radiological Society of North America; IDH: isocitrate dehydrogenase; ACM: Association for Computing Machinery; IEEE: Institute of Electrical and Electronics Engineers; PACS: Picture Archiving and Communication System; GRADE: Grading of Recommendations Assessment, Development, and Evaluations; MENA: Middle East and North Africa; GCC: Gulf Cooperation Council

References

- Singh Y, Hathaway QA, Keishing V, Salehi S, Wei Y, Horvat N, Vera-Garcia DV, et al. Beyond post hoc explanations: a comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. *Bioengineering* (Basel). 2025;12(8):879. [doi](#) [pubmed](#)
- Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. [doi](#) [pubmed](#)
- Barredo Arrieta A, Diaz-Rodriguez N, Del Ser J, Benetton A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
- Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*. 2016:1135-1144.
- Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765-4774.
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Proc IEEE Int Conf Comput Vis*. 2017:618-626.
- Panboonyuen T. Seeing isn't always believing: analysis of Grad-CAM faithfulness and localization reliability in lung cancer CT classification. *arXiv preprint*. 2026.
- Kinger SS, Kulkarni V. A review of explainable AI in medical imaging: implications and applications. *Int J Comput Appl*. 2024;46(11):983-997.
- Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint*. 2013.
- Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. *arXiv preprint*. 2014.
- Smilkov D, Thorat N, Kim B, Viegas F, Wattenberg M. SmoothGrad: removing noise by adding noise. *arXiv preprint*. 2017.
- Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *Proc 34th Int Conf Mach Learn*. 2017:3319-3328.
- Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. *Proc 34th Int Conf Mach Learn*. 2017:3145-3153.
- Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. 2018 IEEE Winter Conf Appl Comput Vis. 2018:839-847.
- Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Adv Neural Inf Process Syst*. 2018;31:9505-9515.
- Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Comput Surv*. 2023;55(13s):1-42.
- Koh PW, Nguyen T, Tang YS, Mussmann S, Pierson E, Kim B, et al. Concept bottleneck models. *Proc 37th Int Conf Mach Learn*. 2020:5338-5348.
- Lucieri A, Bajwa MN, Braun SA, Malik MI, Dengel A, Ahmed S. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. 2020 Int Joint Conf Neural Netw. 2020:1-10.
- Patricio C, Neves JC, Teixeira LF. Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit Workshops*. 2023:3799-3808.
- Patricio C, Teixeira LF, Neves JC. Towards concept-based interpretability of skin lesion diagnosis using vision-language models. 2024 IEEE Int Symp Biomed Imaging. 2024:1-5.
- Jalaboi R, Faye F, Orbes-Arteaga M, Jorgensen D, Winther O, Galimzianova A. DermX: An end-to-end framework for explainable automated dermatological diagnosis. *Med Image Anal*. 2023;83:102647. [doi](#) [pubmed](#)
- Kaczmarzyk JR, Saltz JH, Koo PK. Explainable AI for computational pathology identifies model limitations and tissue biomarkers. *ArXiv*. 2024. [pubmed](#)
- Fontes M, De Almeida JDS, Cunha A. Application of example-based explainable artificial intelligence (XAI) for analysis and interpretation of medical imaging: a systematic review. *IEEE Access*. 2024;12:26419-26427.
- Djoumessi K, Huang Z, Kuhlewein L, Rickmann A, Simon N, Koch LM, Berens P. An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy. *PLOS Digit Health*. 2025;4(5):e0000831. [doi](#) [pubmed](#)
- Quelleg G, Al Hajj H, Lamard M, Conze PH, Massin P, Cochener B. ExplAI: Explanatory artificial intelligence for diabetic retinopathy diagnosis. *Med Image Anal*. 2021;72:102118. [doi](#) [pubmed](#)
- Groen AM, Kraan R, Amirkhan SF, Daams JG, Maas M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? *Eur J Radiol*. 2022;157:110592. [doi](#) [pubmed](#)
- Borys K, Schmitt YA, Nauta M, Seifert C, Kramer N, Friedrich CM, Nensa F. Explainable AI in medical imaging: An overview for clinical practitioners - Beyond saliency-based XAI approaches. *Eur J Radiol*. 2023;162:110786. [doi](#) [pubmed](#)
- Gefter WB, Prokop M, Seo JB, Raoof S, Langlotz CP,

- Hatabu H. Human-AI symbiosis: a path forward to improve chest radiography and the role of radiologists in patient care. *Radiology*. 2024;310(1):e232778. [doi pubmed](#)
29. Plass M, Kargl M, Kiehl TR, Regitnig P, Geissler C, Evans T, Zerbe N, et al. Explainability and causability in digital pathology. *J Pathol Clin Res*. 2023;9(4):251-260. [doi pubmed](#)
 30. Chanda T, Hauser K, Hobelsberger S, Bucher TC, Garcia CN, Wies C, Kittler H, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nat Commun*. 2024;15(1):524. [doi pubmed](#)
 31. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. [doi pubmed](#)
 32. Codella NCF, Nguyen QB, Pankanti S, Gutman DA, Helba B, Halpern AC, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM J Res Dev*. 2017;61(4/5):5:1-5:15.
 33. Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors*. 2004;46(1):50-80. [doi pubmed](#)
 34. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. *Hum Factors*. 2015;57(3):407-434. [doi pubmed](#)
 35. Vereschak O, Bailly G, Caramiaux B. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proc ACM Hum Comput Interact*. 2021;5(CSCW2):1-39.
 36. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. *Int J Hum Comput Stud*. 2023;169:102941.
 37. Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Med Image Anal*. 2023;84:102684. [doi pubmed](#)
 38. Koutsoubis N, Waqas A, Yilmaz Y, Ramachandran RP, Schabath M, Rasool G. Future-proofing medical imaging with privacy-preserving federated learning and uncertainty quantification: a review. *arXiv preprint*. 2024.
 39. Chaddad A, Lu Q, Li J, Katib Y, Kateb R, Tanougast C, et al. Explainable, domain-adaptive, and federated artificial intelligence in medicine. *IEEE/CAA J Autom Sin*. 2023;10(4):859-876.
 40. Jahan S, Adib MRS, Huda SM, Rahman MS, Kaiser MS, Hosen ASMS, Ghimire D, Park MJ. Federated explainable AI-based Alzheimer's disease prediction with multi-modal data. *IEEE Access*. 2025;13:43435-43454.