Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264319

ARTIFICIAL INTELLIGENCE AND INFORMATICS

REVIEW

Explainable artificial intelligence in medical imaging: how to interpret, evaluate, and use artificial intelligence explanations

Gorkem Durak[1]
Halil Ertugrul Aktas[1]
Tugba Akinci D'Antonoli[2,3]
Ulas Bagci[1]

[1]Northwestern University, Machine and Hybrid Intelligence Lab, Department of Radiology, Chicago, United States of America

[2]University Hospital Basel, Division of Diagnostic and Interventional Neuroradiology, Department of Radiology, Basel, Switzerland

[3]University Children's Hospital Basel, Department of Pediatric Radiology, Basel, Switzerland

ABSTRACT

Most artificial intelligence (AI) models used in radiology are black boxes—they produce predictions without explaining the basis of their outputs, raising concerns about clinical safety, accountability, and trust. To address this, a growing body of methods has been developed to help clinicians understand and evaluate AI predictions. This field, known as explainable AI (XAI), aims to help clinicians interrogate, interpret, and critically evaluate AI predictions by identifying factors associated with model outputs. In this educational and practical review, we provide an accessible overview of XAI tailored for practicing radiologists and physicians. We cover the major categories of explanation methods, including saliency maps, perturbation-based and feature-attribution approaches, concept-based methods, and example-based reasoning, as well as uncertainty quantification as a complementary approach for assessing prediction reliability, along with common misconceptions and emerging regulatory obligations. We aim to make XAI easier for healthcare professionals to understand, as effective oversight of AI tools has become a core competency for the modern radiologist.

KEYWORDS

Explainable artificial intelligence, radiology, medical imaging, interpretability, deep learning, artificial intelligence

1. Introduction

Radiology is among the medical specialties most radically transformed by artificial intelligence (AI). Advances in deep learning, a subset of AI based on multilayered neural networks, have enabled algorithms to identify complex patterns in medical images with remarkable speed and, in some tasks, to match or exceed the performance of expert physicians.[1,2] Some of these capabilities have already moved from research to clinical practice, and AI applications are increasingly used to assist radiologists with tasks such as reading mammograms, detecting pulmonary nodules, flagging intracranial hemorrhages, and characterizing thyroid nodules.[3,4] Despite these impressive advances, many AI models remain black boxes, with decision-making processes that are difficult to interpret. They accept an image as input and produce a prediction as output without explaining the basis for their outputs. This opacity creates a cascade of practical problems.[5] Clinicians may not always understand the rationale for a decision or the conditions under which a model is likely to fail. Patients may be hesitant to accept AI-assisted decisions because of a lack of transparency. Researchers and regulators may fail to recognize model failures.

Explainable AI (XAI) is the field dedicated to addressing this problem. It encompasses a diverse set of methods for generating explanations of AI decisions, thereby providing insight into model behavior.[6] Interest in XAI has grown substantially over the past decade, and explainability is now considered not simply a scientific curiosity but a clinical necessity.[7,8] Existing well-written reviews have mapped this terrain from complementary angles: Borys et al. cataloged non-saliency XAI methods by their output representations for a cross-disciplinary readership,[9] and Haupt and Maurer[10] synthesized XAI applications across radiological subspe-

Corresponding author: Gorkem Durak

E-mail: gorkem.durak@northwestern.edu

Received 17 July 2026; revision requested 29 July 2026; last revision requested 05 August 2026; accepted 06 August 2026.

Epub: 21.08.2026

DOI: 10.4274/dir.2026.264319

You may cite this article as: Durak G, Aktas HE, Akinci D'Antonoli T, Bagci U. Explainable artificial intelligence in medical imaging: how to interpret, evaluate, and use artificial intelligence explanations. *Diagn Interv Radiol*. 21 August 2026 DOI: 10.4274/dir.2026.264319 [Epub Ahead of Print].

cialties, along with their methodological limitations and future directions. Both are primarily oriented toward describing which XAI methods exist and how they perform. Building on these contributions, the present review shifts its focus directly to the clinician's standpoint on how AI explanations should be evaluated and clinically supervised. Specifically, it asks how a practicing radiologist should interrogate, appraise, and act on the explanations an AI system generates, including when to distrust a convincing readout.

This educational and practical review is intended for physicians, particularly radiologists, who encounter AI tools in their daily practice or are considering integrating them into their workflow. No background in programming or data science is required to engage with this review, though additional articles are available for those with more technical expertise.[11] This article uses "explainable AI" as an umbrella term encompassing both post-hoc explanation methods and inherently interpretable models. The content of this article is intended to be task-agnostic and broadly applicable to AI applications in detection, segmentation, classification, and other medical imaging tasks, where appropriate. This review has three objectives: first, to explain why explainability matters from a clinical standpoint; second, to introduce the core concepts of XAI in accessible terms; and third, to guide readers in critically evaluating the explanations provided by XAI systems.

Main points

- The black-box nature of artificial intelligence (AI) poses real clinical risks, including the risk that models achieve high accuracy for the wrong reasons.
- Explainable AI (XAI) encompasses methods that help clinicians interpret and critically evaluate AI predictions by providing insight into model behavior.
- No single XAI method is universally superior; the right choice depends on the clinical question, model type, and intended audience.
- Standardized evaluation of XAI quality is currently lacking, creating a critical gap that the field must address.
- XAI is relevant to every physician who uses or oversees AI tools, not only to developers and data scientists. It is increasingly recognized as an important component of safe clinical deployment.

2. Critical questions and answers

2.1 Why is explainability critical for artificial intelligence in medical imaging?

Consider a pathology report that simply states, "Malignant, confidence: 94%." Such a report would provide insufficient supporting information for many high-stakes clinical decisions. Similarly, an AI prediction presented without interpretable supporting information may be difficult for clinicians to evaluate and appropriately contextualize.

The need for explainability in medical AI stems from several distinct yet interconnected requirements:

• **Patient safety:** A model that provides correct answers for the wrong reasons may fail catastrophically for patients who differ even slightly from its training population. Explanations help clinicians assess whether the model's behavior aligns with established medical knowledge before acting on its output.

• **Error detection:** When an AI tool produces an unexpected or incorrect result, explainability can help investigate the error by identifying potential sources of model failure.

• **Trust calibration:** Clinicians should neither blindly trust AI recommendations nor reflexively dismiss them. Explanations enable the calibrated trust needed for safe human–AI collaboration.

• **Transparency and oversight:** Regulatory authorities emphasize information-sharing and transparency in defined circumstances.[12,13] These include providing meaningful information about the logic, the significance of the processing, and its anticipated consequences, without necessarily disclosing a model's full internal mechanics or requiring the use of a specific XAI method. Further regulations will be discussed in Section 5.

• **Professional accountability:** Radiologists remain professionally accountable for clinical judgments within their scope of practice, but legal liability for AI-related harm may be shared among clinicians, healthcare institutions, manufacturers, and other actors, depending on the jurisdiction and circumstances.

2.2 Do I need to learn about explainable artificial intelligence as a clinician?

A common misconception is that XAI is exclusively a technical concern, relevant only to the computer scientists and data engineers who build AI models, not to the clinicians who use them. This view is mistaken and potentially dangerous.

There are at least three roles in which a physician engages directly with XAI:

• **As users of AI tools,** radiologists and other physicians increasingly rely on AI-assisted platforms in daily clinical practice. When a tool provides an uncertainty estimate alongside its prediction, the clinician needs to understand what that estimate means, its limitations, and how much weight to give it.

• **As critical evaluators,** physicians are responsible for assessing any new AI tool proposed for adoption. Critically examining the explanations a model provides, not just its numerical performance, is essential to that evaluation.

• **As feedback providers,** clinicians are uniquely positioned to identify when an AI model relies on anatomically implausible or clinically inappropriate features or behaves inconsistently with established disease mechanisms. This feedback is invaluable for improving the model. Radiologists who understand XAI can communicate these concerns in a way that enables developers to act.

Just as pilots rely on cockpit instruments while remaining mindful of their limitations, clinicians should use XAI to monitor AI predictions while recognizing that explanations may themselves be imperfect. XAI is therefore a tool for oversight, not proof that a model's prediction is correct.

2.3 Is a high-accuracy model with a convincing explanation enough to be trusted?

Performance metrics assess outcomes, not reasoning. Metrics such as area under the curve (AUC), sensitivity, and specificity indicate how often a model makes correct predictions, but they do not reveal whether those predictions are grounded in clinically valid reasoning. This distinction is not purely academic. Deep learning models on chest radiographs have repeatedly been shown to exploit shortcuts rather than pathology: pneumonia and COVID-19 models rely on portable imaging and institution-specific acquisition characteristics, whereas pneumothorax models rely on the chest drains used to treat the condition. Because these cues reflect clinical workflow rather than biology, they inflate apparent performance and fail when the shortcut is absent.[14-16]

In mammography, models trained on images from a single manufacturer have learned vendor-specific image characteristics that correlate with the training labels, degrading performance when deployed on equipment from other vendors.[17] This phenomenon, in which a model learns statistically convenient but clinically irrelevant correlations, is known as shortcut learning or the Clever Hans effect. The name refers to the famous horse that appeared to solve arithmetic problems but was actually reading its trainer's cues.[18]

XAI provides tools to detect it: a model that reaches the correct answer while focusing on the wrong part of the image should raise an immediate red flag, regardless of its overall accuracy. Model accuracy and explanation quality are therefore distinct properties.[19] Explanation quality also has two dimensions that are often conflated. Plausibility is the degree to which an explanation aligns with clinical expectations; for example, a heatmap highlighting the lesion the radiologist would have chosen is plausible. Faithfulness is the degree to which an explanation reflects the computation the model actually performed (Figure 1). Plausibility is judged by the clinician; faithfulness is assessed by comparing the explanation with the model's actual decision process. The two can diverge in either direction. When a model works as intended but relies on shortcut learning, scanner artifacts, or spurious correlations, an explanation would be faithful yet implausible. Conversely, an explanation may highlight the clinically relevant region while reflecting little of what the model learned: plausible but unfaithful, and therefore not valid evidence of the model's decision basis. This second scenario is arguably more concerning because, by inspection, it is indistinguishable from a genuinely faithful explanation. Accordingly, an explanation should not be regarded as proof of validity and should be interpreted cautiously.

The same skepticism applies when different XAI tools disagree, which happens more often than clinicians expect. Because methods define importance differently [Gradient-weighted Class Activation Mapping (Grad-CAM) follows gradients, Local Interpretable Model-agnostic Explanations (LIME) masks regions, and SHapley Additive exPlanations (SHAP) quantifies feature contributions], minor differences between explanation methods are expected, but their clinical significance should be interpreted in context. Large discrepancies, such as one method highlighting a pulmonary mass while another highlights the image periphery, raise concerns about explanation instability, method dependence, or reliance on complex model behavior and warrant further evaluation. The wrong response is to choose the output that matches clinical intuition, which introduces confirmation bias and undermines XAI's purpose as an auditing tool. Instead, consult quantitative validation data, raise the discrepancy with the vendor or informatics team, and apply independent clinical judgment without deferring to either output.

3. Workflow of explainable artificial intelligence in medical imaging

Understanding how XAI fits into the broader AI-assisted radiology pipeline helps clinicians clarify the origin of explanations and what they represent. The overall workflow, illustrated in Figure 2, consists of three interconnected phases.

3.1 Image acquisition and preprocessing

The workflow begins with image acquisition, regardless of the image [computed tomography (CT), magnetic resonance imaging, radiography, or ultrasound], which serves as input to the AI system. Before entering the AI model, images typically undergo preprocessing: standardizing pixel values, resizing to a fixed dimension, and, sometimes, normalizing intensity. Software typically handles these steps automatically, but they matter because they can affect what the model sees and, consequently, what an explanation highlights.

One important practical consideration is distribution shift: if a model is trained on images from one scanner type or acquisition protocol and then deployed on images from another, both the model's performance and the quality of its explanations may degrade. This is one reason why explanations that seem clinically sensible on validation data may appear unexpected in real-world practice.

3.2 Model inference and explainability strategy

After preprocessing, the image is fed into the AI model. The model then processes the image through a series of mathematical transformations across multiple layers, ultimately producing an output such as a classification label ("pneumonia/no pneumonia") or a probability score ("82% likelihood of malignancy").

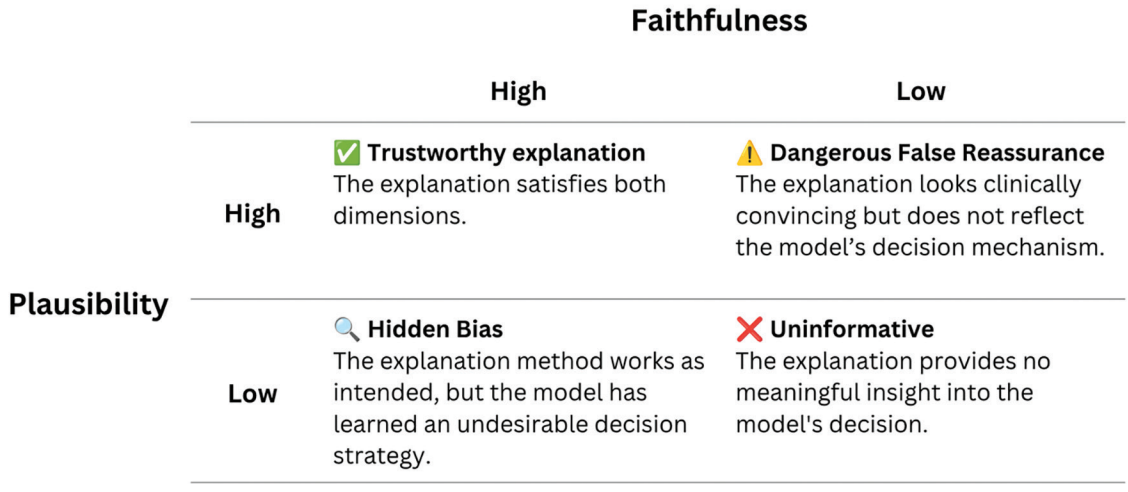

Figure 1. Plausibility and faithfulness are distinct dimensions of an explanation. Plausibility assesses whether an explanation aligns with clinical expectations and is evaluated by the clinician. Faithfulness assesses whether the explanation accurately represents the model's underlying computation and must be evaluated against the model itself.

The explainability strategy refers to the method used to generate an explanation for the model's output. Broadly, explanations can be generated in two ways (Figure 3):[20]

• **Ante-hoc (intrinsic) methods:** Ante-hoc models are designed so that their predictions are inherently interpretable, eliminating the need for separate post-hoc explanation methods. Although these models have traditionally been thought to sacrifice performance for transparency, this trade-off is increasingly contested. A growing body of work argues that interpretability should be incorporated during model development rather than retrofitted after prediction.[21]

• **Post-hoc methods:** These techniques are applied after training to generate explanations for a black-box model's predictions. Many AI tools currently deployed in radiology rely on complex black-box models, which is why post-hoc methods dominate clinical XAI in practice. Post-hoc methods include saliency maps, perturbation-based methods, and feature attribution scores, all described in Section 4.

3.3 Clinical interpretation and feedback

The final phase, the one most directly relevant to the practicing physician, is the clinical interpretation of the explanation. This step is not passive. A thoughtful radiologist should ask: Does this explanation make sense? Does the model attend to the correct

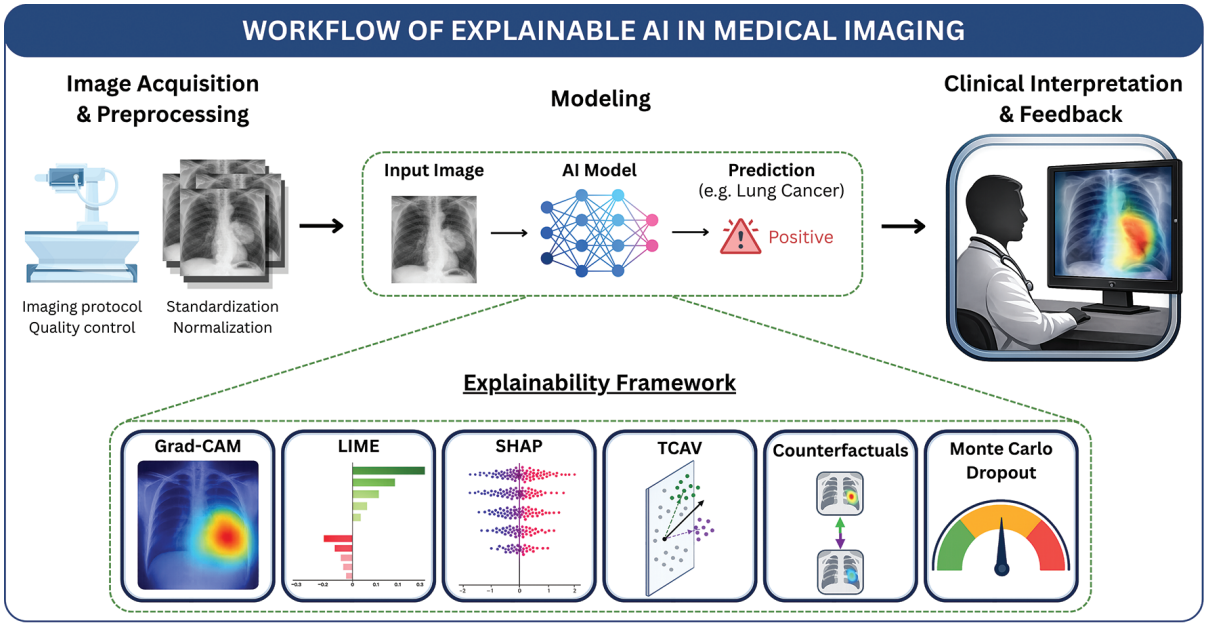

Figure 2. Overview of the explainable artificial intelligence (AI) workflow that illustrates how explanation methods provide complementary information to interrogate AI predictions during clinical oversight. Grad-CAM, gradient-weighted class activation mapping; LIME, local interpretable model-agnostic explanations; SHAP, SHapley Additive exPlanations; TCAV, testing with concept activation vectors.

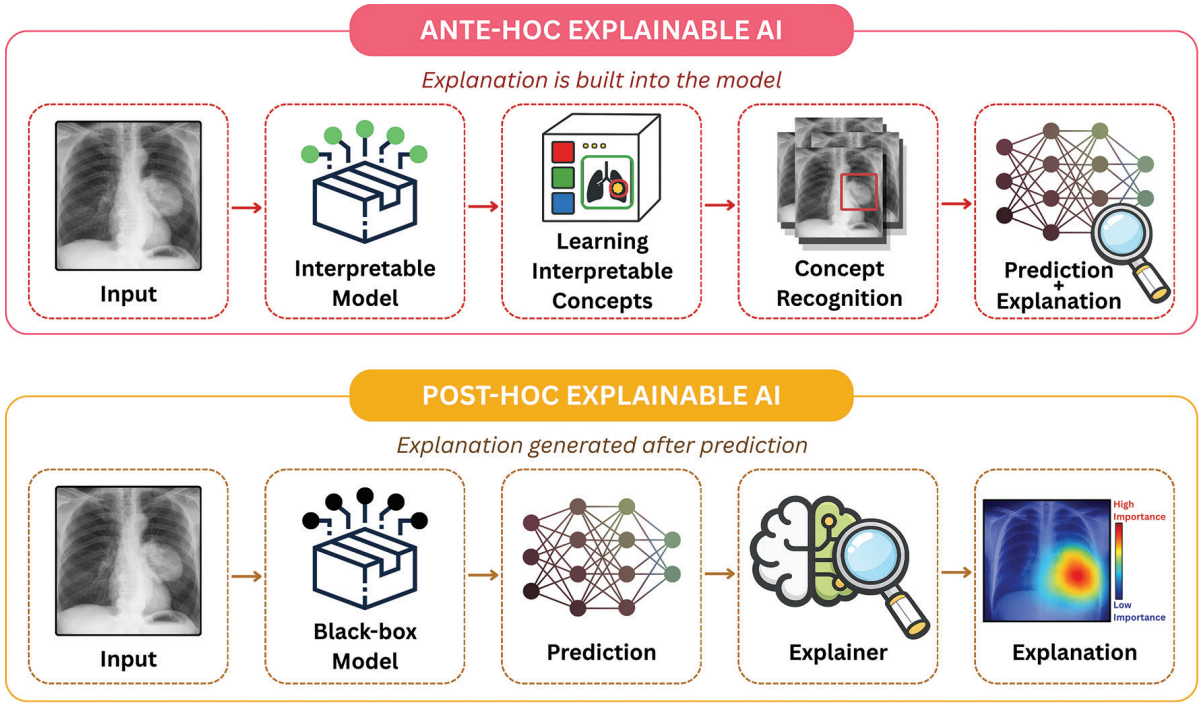

Figure 3. Representative comparison of post-hoc and ante-hoc explainable artificial intelligence approaches.

region, or is it focusing on irrelevant areas (such as labels, artifacts, or patient identifiers)? Does the model's confidence match the case's difficulty?

Quality assurance for AI models requires ongoing, critical collaboration between radiologists and developers. When clinicians routinely review explanation outputs alongside model performance, they may detect implausible behavior, shortcut learning, or unexpected failures that aggregate performance metrics alone might miss. Explainability can strengthen this cycle by helping clinicians and developers identify otherwise hidden model limitations and failure modes, enabling radiologist feedback, model refinement, and iterative improvement of AI systems.

4. Categories of explainable artificial intelligence in medical imaging

Researchers have developed and applied a wide variety of XAI methods in medical imaging. For educational purposes, these methods are grouped into six broad categories, with perturbation-based and feature-attribution methods discussed together in Section 4.2. These categories are not mutually exclusive, and many XAI methods can be described along multiple dimensions (for example, how explanations are generated, what they explain, or how they are presented) (Figure 4). Most of these categories address what the model is attending to or why it produced a given output, using spatial, feature-level, concept-level, or example-level accounts. Uncertainty quantification does not explain why a prediction was made but complements explainability by estimating the reliability of that prediction. Table 1 provides a comparative summary.

4.1 Visual and spatial explanation methods

The most widely used XAI method in radiology is the saliency map, a visual overlay typically rendered as a heatmap that highlights the image regions most important to the model's prediction. When a radiologist opens an AI platform and sees a warm-colored region superimposed on a chest CT indicating where the algorithm detected a nodule, they are viewing a saliency map.

One of the most commonly used methods for generating saliency maps is Grad-CAM.[22] It traces gradients flowing backward from the predicted class to the final convolutional layer.

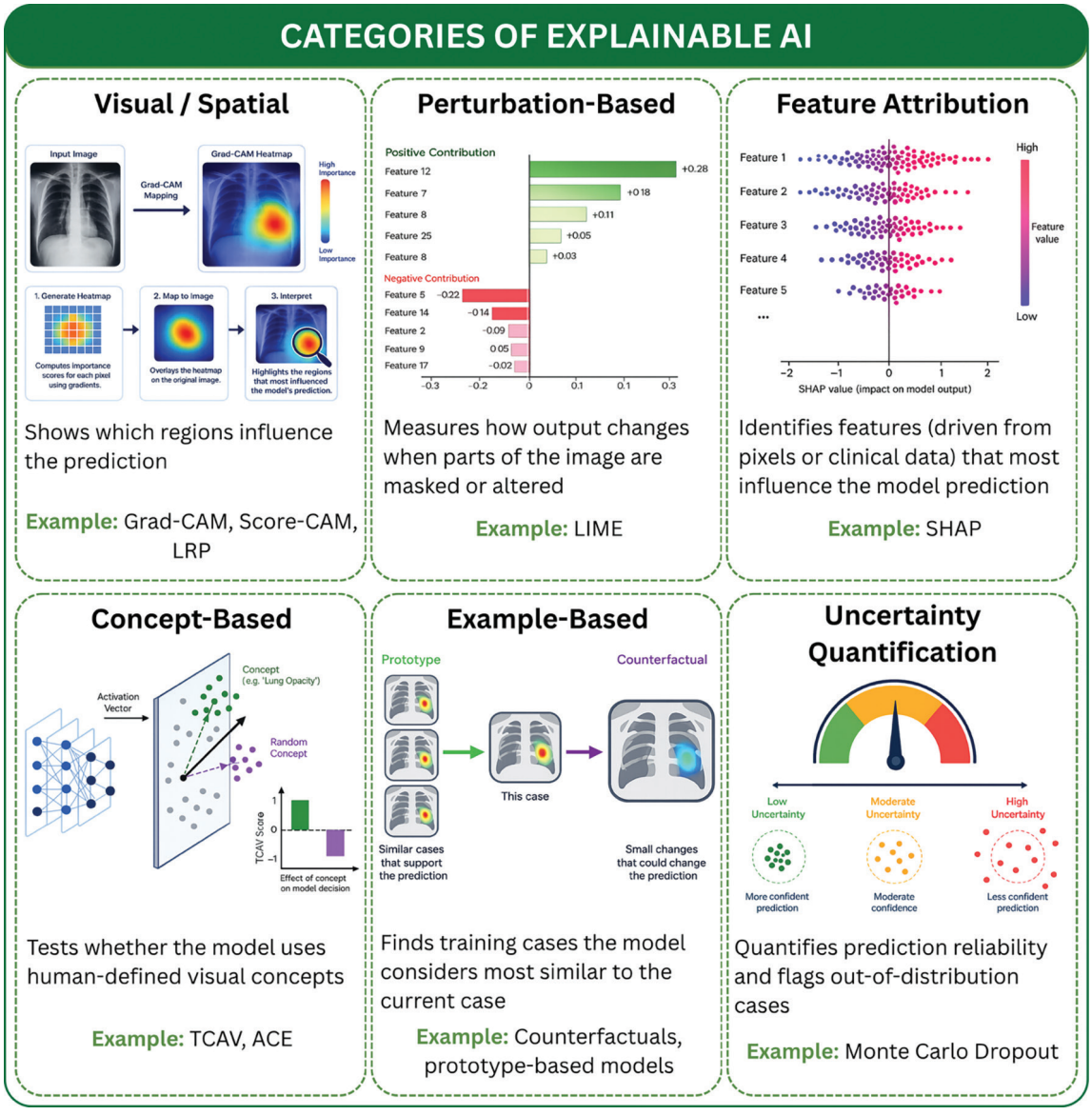

Figure 4. Overview of the major categories of explainable artificial intelligence methods with representative illustrations. Grad-CAM, gradient-weighted class activation mapping; LIME, local interpretable model-agnostic explanations; SHAP, SHapley Additive exPlanations; TCAV, testing with concept activation vectors; ACE, automated concept-based explanation.

Table 1. Summary of explainable artificial intelligence categories in medical imaging

Category	What it shows	Typical example	Strengths	Weaknesses
Visual/spatial	Which image regions influenced the prediction	Grad-CAM saliency map	✓ Intuitive and easy to understand; visually appealing for clinicians ✓ Computationally fast; widely implemented	✗ Spatial only—does not explain why a region matters ✗ Maps can reflect image structure rather than the model's decision process
Perturbation-based	How the output changes when parts of the image are masked	LIME	✓ Model-agnostic—works with any architecture ✓ Interpretable to non-technical audiences	✗ Can be computationally expensive—many forward passes required ✗ Local approximations may not reflect global model behavior ✗ Sensitive to masking strategy
Feature attribution	Which features (numerical or image-derived) contribute to the prediction	SHAP	✓ Quantifies feature contributions ✓ Local attributions can be aggregated for cohort-level model auditing ✓ Useful for multivariable clinical models	✗ Many implementations make simplifying assumptions about feature dependence, which may distort attributions for correlated variables ✗ Can be computationally expensive at scale
Concept-based	Whether the model uses human-defined visual concepts (e.g., spiculation, mass shape)	TCAV, ACE	✓ Speaks the language of radiology—spiculation, margin, and shape ✓ Enables direct audit against clinical reporting criteria	✗ Requires pre-defined, expertly labeled concept sets ✗ Concept quality directly limits explanation validity ✗ Difficult to scale across diseases
Example- and case-based	How the model relates the current case to representative examples	Counterfactuals, prototype matching	✓ Mirrors the case-based reasoning familiar to clinicians ✓ Provides concrete examples to support interpretation	✗ Prototype and retrieval methods depend on the representativeness of the reference library ✗ Counterfactual explanations require validation of anatomical realism, clinical feasibility, and clinical plausibility
Uncertainty quantification	How reliable the prediction is and whether the case may be out of distribution	Monte Carlo dropout, deep ensembles	✓ May indicate atypical or borderline cases before a wrong decision is made ✓ Supports other XAI methods and safe human oversight	✗ Uncertainty estimates themselves may be miscalibrated in novel contexts ✗ Does not explain why the model made its prediction—only how reliable it is

Grad-CAM, gradient-weighted class activation mapping; LIME, local interpretable model-agnostic explanations; SHAP, SHapley Additive exPlanations; TCAV, testing with concept activation vectors; ACE, automated concept-based explanations; XAI, explainable artificial intelligence.

In simpler terms, it estimates which image regions contributed most strongly to the prediction. The result is a coarse localization map that can be overlaid on the original image.

Several Grad-CAM variants have been proposed. Grad-CAM++ extends Grad-CAM to better capture cases where multiple instances of the same class appear in a single image.[23] Eigen-CAM uses a different mathematical approach that does not rely on class-specific gradients.[24] Score-CAM avoids gradients altogether, instead using activation maps to mask the input image and scoring the resulting change in the model's output.[25] All of these variants aim to provide a spatially meaningful explanation of the model's prediction.

Other backpropagation-based methods include Layer-wise Relevance Propagation,[26] which traces relevance scores backward from the output through every layer of the network to the input pixels, and Integrated Gradients,[27] which satisfies the completeness property, meaning that the sum of the attribution scores equals the difference between the model's output for the input image and its output for a baseline reference image.

What radiologists should know: Saliency maps are intuitive and visually compelling, but visual plausibility is not evidence of validity, as discussed in Section 5.1. The reliability of any given map depends on the method and the specific case.

4.2 Perturbation-based and feature attribution methods

Beyond saliency maps, explainability techniques can also identify which inputs or features contribute most strongly to a model's predictions. Although perturbation-based and feature-attribution methods represent different conceptual approaches, they are often discussed together because they provide complementary insights. LIME[28] is a widely used perturbation-based feature-attribution method. It is fundamentally a local explanation method that clarifies individual predictions and is most commonly used for case-level interpretation. It divides an image into superpixels and generates multiple modified versions by selectively masking these regions. It then fits an interpretable surrogate model to approximate the original model's behavior, highlighting the regions or features that contributed most to a specific prediction. Because LIME is model-agnostic, it can be applied to a wide range of AI systems, though the resulting explanations may be sensitive to parameter choices and the method can be computationally intensive.

SHAP,[29] derived from cooperative game theory, assigns each feature a contribution value that reflects its influence on the prediction. Like LIME, SHAP is fundamentally a local explanation method that clarifies individual predictions. However, its local feature attributions can be aggregated across patient cohorts, making SHAP particularly well suited

for global feature attribution. Conceptually, it estimates how much each feature influences the model's prediction across different feature combinations and assigns each feature a "fair share" of the prediction. Aggregating these contributions across multiple cases enables SHAP to identify variables that consistently influence model predictions across a population.

What radiologists should know: LIME and SHAP offer complementary forms of explainability—LIME is more commonly used for individual cases, whereas SHAP is more commonly used across populations. Both methods can be applied to multimodal AI models that combine imaging with electronic health record data, laboratory values, demographic information, or genomic markers, helping clinicians understand which inputs contribute most to the final prediction. They may also reveal potential biases when models rely heavily on confounding variables rather than disease-related characteristics.

4.3 Concept-based explainability

One of the most clinically intuitive categories of XAI is the concept-based explanation. Rather than asking "which pixels matter?", concept-based methods ask, "does this model's behavior correlate with human-defined visual concepts?"—for example, spiculation of a lung nodule, heterogeneity of a liver lesion, or cortical disruption of a bone.

Testing with concept activation vectors (TCAV)[30] is a prominent method in this category. It first defines a clinically meaningful concept using representative examples, contrasts these examples with an appropriate control set, and then fits a linear classifier whose decision boundary identifies the concept's direction in the model's internal feature space (for example, images with spiculated vs. smooth nodule margins). The resulting concept activation vector encodes the concept in the neural network's mathematical space. Then, TCAV estimates how sensitive the model's predictions are to the presence of that concept, formally expressed as the fraction of cases in which the concept increases the model's score for a specified target class.

Automated concept-based explanations (ACE)[31] extend the TCAV idea by automating concept discovery, reducing, though not entirely eliminating, the need for manually labeled concept datasets since the discovered concepts still require expert review for clinical interpretation. ACE generates image segments and clusters their internal feature representations to identify groups of visually similar regions that appear consistently across many images within the same class, treating these clusters as automatically discovered concepts.

Concept-based methods are especially valuable in radiology because radiologists already think in visual terms. A model that can be interrogated with the question "Is this model using the pleural effusion to predict malignancy?" is far more clinically useful than one that simply produces a diffuse heatmap. As these methods mature, they hold great promise for aligning AI reasoning with established radiological reporting frameworks.

What radiologists should know: Concept-based methods are among the most clinically aligned, interrogating models by using terms such as spiculation or cortical disruption rather than pixel gradients. When evaluating an AI tool, assess it against the specific imaging criteria already used in the relevant specialty. Otherwise, its explanations may be technically valid yet clinically uninformative.

4.4 Example-based and case-based explanations

Example-based methods explain a model's prediction by referring to representative examples, retrieved reference cases, or learned prototypes rather than to abstract features or pixel maps. This approach mirrors the pattern-recognition process familiar to radiologists, who often interpret imaging findings by comparing them with previously encountered cases.

Counterfactual explanations address this question: "What would the image need to look like for the model to predict a different outcome?" The concept originated in automated decision-making under data-protection law[32] and has since been extended to imaging tasks, where it may overlap with example-based reasoning when reference images or prototypes are used. For example, given a mammogram classified as high risk, a counterfactual explanation might show what the mammogram would need to look like, perhaps with a smaller or smoother lesion, to be classified as low-risk. This contrastive structure aligns with human reasoning and may improve clinicians' understanding of AI decisions.

Prototype-based models[33] go a step further by incorporating example-based reasoning directly into the model architecture. The model learns a set of representative image prototypes in its internal feature space for each class. For any new image, the model's prediction is based on similarity to these prototypes, and the explanation takes this form: "This region of the new image resembles this prototype, which is associated with malignancy." However, this requires careful design to avoid misleading or degenerate prototypes.

What radiologists should know: Example-based explanations are also intuitive for clinicians because they mirror case-based reasoning, but their trustworthiness depends on the quality, realism, and clinical validity of the generated or retrieved examples. For retrieval-based systems, assess whether the reference cases represent the relevant patient population. For prototype-based models, examine how the learned prototypes are constructed and validated.

4.5 Uncertainty quantification

An often-overlooked yet critically important aspect of AI interpretation in clinical practice is uncertainty quantification. Most deep learning models produce a single probability score, which does not directly reflect the prediction's reliability. The probability score indicates how strongly the model favors a prediction, whereas uncertainty quantification estimates its reliability. A model may assign a high probability to an incorrect prediction, making such failures difficult to recognize. Uncertainty quantification methods address this limitation by estimating how much to trust a prediction.[34]

Uncertainty estimation has several important clinical applications. It can help identify difficult or ambiguous cases, highlight predictions that may warrant additional scrutiny, and flag out-of-distribution cases. These images differ substantially from the model's training data, and predictions can be less reliable.[35] In practice, high-uncertainty cases can be automatically flagged for human review, an approach increasingly advocated for safe AI deployment.

Several techniques can estimate a model's uncertainty. Monte Carlo Dropout is among the most widely used. It repeatedly evaluates the same image with different dropout patterns during inference, and the variability among the resulting predictions serves as the uncertainty estimate.[36] Deep ensembles, instead, compare the predictions of several independently trained models.[37] Both methods are primarily used to estimate epistemic uncertainty, which reflects uncertainty due to limited model knowledge

or unfamiliar inputs. By contrast, aleatoric uncertainty reflects inherent ambiguity or noise in the input data and typically requires a model explicitly designed to estimate it. Such estimates complement explainability methods such as saliency and attribution maps: whereas those maps indicate where the model looked or which features it relied on, uncertainty indicates how reliable the model's prediction is.

What radiologists should know: A system that provides well-calibrated, externally validated uncertainty estimates may support safer clinical oversight than one that provides no uncertainty information. Uncertainty estimates help clinicians act appropriately on model outputs, for example, by deferring high-uncertainty cases for immediate expert review. However, users should also be cautious; an AI system may report uncertainty, but it cannot determine whether its estimates remain trustworthy in a novel clinical context.[38]

4.6 How do I choose between explainable artificial intelligence methods for a given task?

No single XAI method is best; the right choice depends on the clinical question and the intended audience. The clinical question should be considered first:

- To understand where the model is focusing, a saliency method such as Grad-CAM is the natural first choice.
- To identify feature importance for individual predictions, LIME would be preferable.
- To identify feature importance across a patient cohort, aggregated SHAP analyses are more appropriate.
- To determine whether the model is using radiologically meaningful concepts, such as spiculation or lesion heterogeneity, concept-based methods, such as TCAV, are the most directly informative.
- To understand which prior cases most influenced the prediction, or what would need to change for the prediction to differ, prototype retrieval or counterfactual methods should be used.
- To assess prediction reliability, uncertainty quantification is the appropriate choice.

Additionally, the audience should be considered. Heatmaps may be particularly comprehensible for radiologists because they preserve the spatial context of imaging findings, whereas SHAP feature plots may facilitate multidisciplinary discussions involving clinical variables. Prototype and counterfactual explanations can provide intuitive examples that nonexpert users can interpret more easily. When needed, validate findings with complementary methods. A practical guide for selecting and critically evaluating XAI methods is provided in Table 2.

5. Limitations, common misconceptions, and future directions

Although XAI is a necessary complement to AI, it has limitations and can be misused. This section reviews common misconceptions, limitations, and future directions in XAI and summarizes the key concepts in Figure 5.

5.1 Visual appeal does not equal true reasoning

Perhaps the most important misconception in clinical XAI is that a visually convincing heatmap constitutes a reliable explanation. A heatmap highlighting the correct region of an image is reassuring, yet it does not provide sufficient evidence of faithful model reasoning or guarantee that it will continue to highlight the correct region in other cases or on other scanners.

Adebayo et al.[39] showed in a landmark study that several widely used saliency methods produced similar heatmaps even after progressively randomizing the model's learned parameters. This finding suggested that the resulting explanations were driven largely by image characteristics such as edges and textures rather than by the model's learned representations. A method that produces the same explanation regardless of what the model has learned cannot be considered a faithful explanation of the model itself (see the "Dangerous False Reassurance" in Figure 1).

Quantitative benchmarking has also shown that saliency maps do not always localize lesions accurately. In a study evaluating Grad-CAM, Grad-CAM++, and Eigen-CAM on mammography,[40] Pointing Game Scores,[41] which measure localization accuracy by checking whether the saliency map's peak activation falls within the expert-annotated lesion region, were only 0.41, 0.30, and 0.35, respectively, substantially lower than the visual impression of reliability would suggest.

Debate continues over whether the entire post-hoc XAI paradigm might be a category error in high-stakes medical settings—if explanations generated after a model reaches its conclusion cannot reliably reconstruct its actual reasoning, then their clinical value, however intuitive they may seem, remains fundamentally limited. Proponents of this view argue that the field should instead prioritize interpretable-by-design approaches, in which transparency is embedded directly in the model architecture rather than inferred from post-hoc explanations.[19,21,42] Post-hoc methods remain crucial for auditing the black-box models that currently dominate clinical deployment, but when an ante-hoc model of comparable performance exists, there is a strong argument for preferring it.

5.2 Risk of overtrust and automation bias

Counterintuitively, explanations accompanying AI predictions may not always reduce inappropriate trust and, in some cases, may even reinforce it.[43] This phenomenon, known as automation bias, is the tendency to over-rely on AI outputs despite conflicting clinical evidence, a risk that may be heightened by visually persuasive explanations. Even a correctly placed heatmap should prompt independent confirmation of the finding rather than deference to the model (see Section 2.3).

Beyond the content of an explanation, the way it is presented, including its on-screen position, the opacity of a heatmap overlay, and the visual prominence of uncertainty indicators, can fundamentally alter how a clinician responds to it, regardless of its accuracy. Emerging human factors research suggests that AI interface design can influence how clinicians interact with AI outputs, making interface design integral to safe AI deployment rather than an afterthought.[44] Cognitive load is another important consideration. Although evidence specific to explainability remains limited, studies on radiologist fatigue underscore that increased cognitive demands can adversely affect diagnostic performance, emphasizing the need for concise, clinically interpretable explanation interfaces.[45]

5.3 Lack of standardized evaluation metrics

In a systematic review of primary diagnostic test accuracy studies published from January 2016 to January 2021, 490 radiology studies used end-to-end deep learning, of which 179 (37%) incorporated XAI. Only one study used a formal measure to assess explanation quality.[46] This gap is striking, given that the field routinely evaluates diagnostic performance using rigorous statistical testing.

One of the most pressing unsolved problems in clinical XAI is the lack of standardized, agreed-upon metrics for evaluating explanation quality.[47] Unlike diagnostic AI performance, where metrics such as AUC, sensitivity, and specificity are

Table 2. Practical guide for selecting and critically evaluating explainable artificial intelligence methods

Clinical question	Recommended XAI method(s)	What should the radiologist verify?	Example
Is the AI focusing on the clinically relevant region?	Grad-CAM or other visual explanation methods	Verify localization accuracy and review evidence for clinical validation. Be aware that visually plausible localization does not necessarily indicate a faithful explanation.	Pulmonary nodule detector highlights the nodule rather than ribs or scanner artifacts.
Why did the AI make this prediction for this patient?	LIME, SHAP	Assess whether the reported feature importance is clinically plausible and review available evidence that the explanation method has been validated for faithfulness.	Identify which imaging and/or clinical features most influenced the prediction for a liver lesion.
Is the model relying on clinically meaningful features or concepts rather than shortcuts?	SHAP, TCAV, Grad-CAM	Determine whether the highlighted concepts represent genuine imaging biomarkers rather than dataset-specific artifacts or shortcuts.	The mammography model relies on spiculated margins rather than acquisition markers.
How reliable is the model's prediction in this case?	Uncertainty quantification	Verify calibration and external performance.	The model flags an equivocal pulmonary embolism study for radiologist review.
How reliable is the prediction in difficult or unusual cases?	Uncertainty quantification	Review the uncertainty estimate and consider whether the case may be out-of-distribution.	A workflow flags an atypical chest radiograph—unlike those in the training data—as high-uncertainty and routes it for priority human review.
Which prior cases drove this prediction, or what would change it?	Prototype methods, counterfactual explanations	Check whether the retrieved prototypes or reference cases resemble your patient population and whether the counterfactual change is clinically meaningful.	A mammogram classified as high risk is shown alongside a counterfactual with a smaller, smoother lesion that would reclassify it as low risk.
How are imaging and clinical variables combined?	SHAP, LIME	Assess how multimodal information contributes to prediction and whether the model's decision basis is clinically plausible.	A hepatocellular carcinoma prognosis model combines imaging and laboratory data.
Should this AI model be trusted before deployment?	Comprehensive XAI evaluation	Review explanation validation, dataset compatibility, workflow integration, limitations, and post-deployment monitoring.	A comprehensive audit of a chest CT AI model is conducted before clinical implementation.

AI, artificial intelligence; Grad-CAM, gradient-weighted class activation mapping; LIME, local interpretable model-agnostic explanations; SHAP, SHapley Additive exPlanations; TCAV, testing with concept activation vectors; XAI, explainable artificial intelligence; CT, computed tomography.

widely understood and routinely reported, no single universally accepted standard exists for evaluating explanation quality. Several evaluation frameworks have been proposed. In addition to the aforementioned Pointing Game Score, insertion and deletion scores assess whether progressively removing or restoring the regions highlighted by an explanation produces the expected change in the model's prediction confidence.[48] Faithfulness metrics also evaluate whether the explanation accurately represents the features the model used to reach its output.[49] Qualitative evaluation using a radiologist's judgment is also valuable but introduces subjectivity and interobserver variability.

The lack of standardization has practical consequences. It means that two AI products can both claim to provide XAI-enhanced outputs without a meaningful way to compare the quality or reliability of their explanations. It also makes the XAI literature difficult to synthesize, as methods are rarely evaluated on the same benchmarks. Therefore, developing and adopting standardized evaluation criteria for XAI in clinical imaging is a top priority for the field.

5.4 Regulatory frameworks and transparency obligations

AI tools in clinical radiology do not deploy in a regulatory vacuum, and the frameworks that govern their use directly shape how transparency, documentation, human oversight, and communication of AI outputs are applied in clinical practice. In the United States, the Food and Drug Administration (FDA) regulates many AI-enabled imaging software products as Software as a Medical Device or as other medical-device software functions, depending on their intended use. Because the statutory non-device clinical decision-support exclusion under the 21st Century Cures Act does not apply to software that acquires, process-

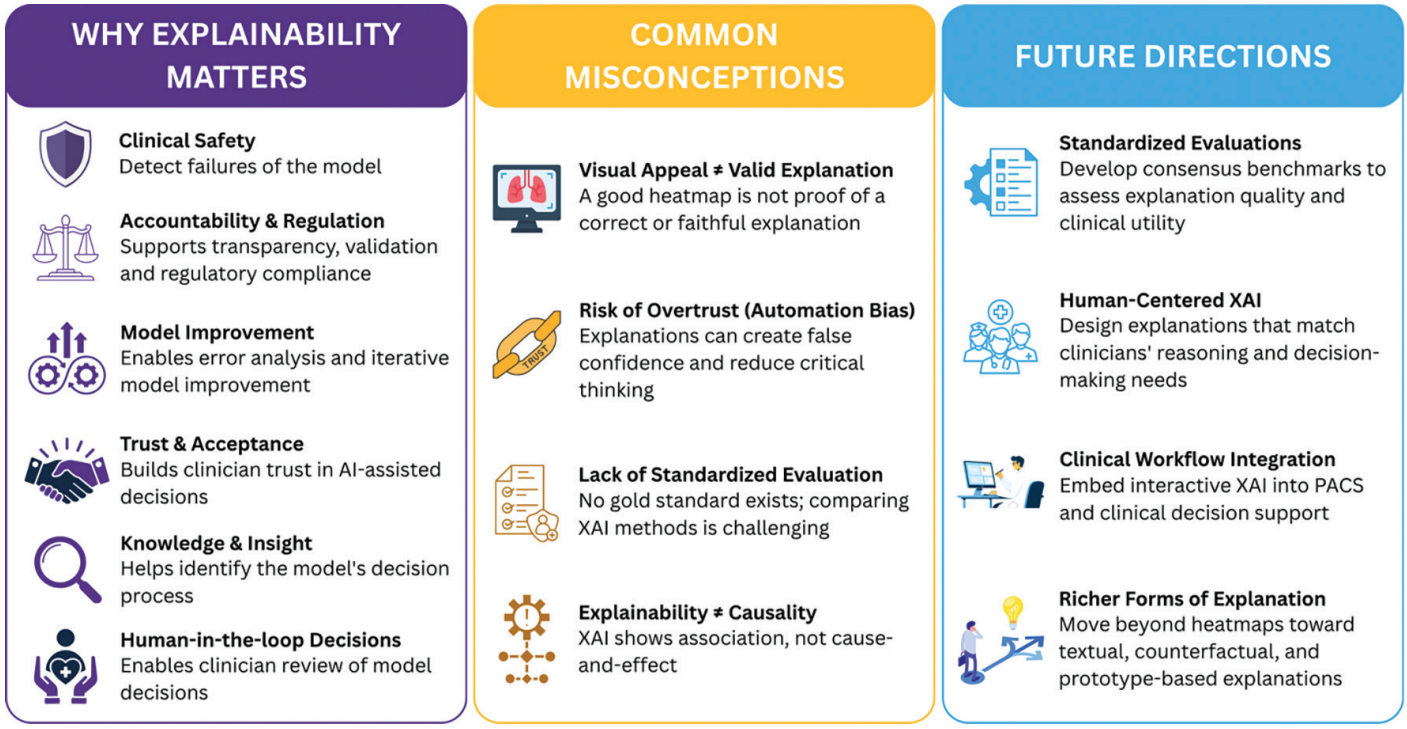

Figure 5. Summary of the clinical importance, common misconceptions, and future directions of explainable artificial intelligence. XAI, explainable artificial intelligence; PACS, Picture Archiving and Communication System.

es, or analyzes a medical image, imaging AI tools fall outside that exclusion and remain subject to FDA device oversight, regardless of the transparency or explainability features they provide. The Good Machine Learning Practice guiding principles,[50] jointly published by the FDA, Health Canada, and the United Kingdom Medicines and Healthcare products Regulatory Agency (MHRA) in 2021, outline 10 principles that encourage developers to provide users with clear information about model performance across subgroups, training data characteristics, and post-deployment monitoring. Building on this foundation, the FDA, Health Canada, and MHRA jointly published Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles in June 2024,[51] which further emphasized the need to communicate information required for safe and informed use, including intended use, performance, limitations, data characteristics, lifecycle monitoring, and, where appropriate, information about the model or algorithmic approach. Although these are nonbinding guiding principles rather than binding regulations, they signal a clear regulatory direction and increasingly shape regulatory submissions and expectations for product transparency.

In Europe, the General Data Protection Regulation (GDPR) imposes transparency obligations on certain forms of automated decision-making, but it should not be read as creating a universal right to a technical explanation for every AI-assisted medical decision.[13,32] More recently, the European Union AI Act entered into force in August 2024 and classified many AI-enabled medical devices as high-risk AI systems, subject to the corresponding regulatory requirements.[52,53] It requires providers to design these systems to be sufficiently transparent so that deployers can interpret their output and use them appropriately and to provide instructions covering logging, human oversight, and the system's limitations and performance characteristics.[54] Notably, the Act's individual right to an explanation of a specific automated decision does not apply to medical devices.[12] This reinforces, as with the GDPR, that there is no general right to a case-level technical explanation for imaging AI.

This Act's timeline has since been amended: Regulation (EU) 2026/1744, which entered into force on 27 July 2026, postponed the high-risk obligations to 2 December 2027 for standalone systems and to 2 August 2028 for AI embedded in regulated products, including medical devices. Accordingly, although many AI-enabled medical devices remain classified as high-risk, the corresponding obligations do not yet apply.

5.5 Future directions

Among the most consequential advances in AI, and by extension in XAI, is the emergence of foundation models. These large multimodal systems are trained on vast datasets of paired images and text and introduce a qualitatively different approach to communicating AI predictions and their supporting rationales. Whereas conventional image classification models often require separate post-hoc explanation methods, vision-language models such as GPT-4 and Med-Gemini can generate natural-language outputs alongside their predictions.[55,56] These outputs create the appearance of transparency, a convincing rationale, and stepwise reasoning but should not be interpreted as revealing the model's underlying computational process.[57] They may instead be post-hoc rationalizations that describe a plausible reasoning pathway without faithfully representing how the prediction was reached. A critical priority for the field will therefore be to develop safeguards against hallucinated or ungrounded natural-language explanations. These generated rationales should be evaluated independently along four complementary dimensions: clinical usefulness, factual accuracy, grounding in the input data, and faithfulness to the underlying prediction process.[58]

Several further developments in XAI for radiology are likely to shape its trajectory over the coming years.

• **Standardization and regulation:** As discussed in detail in the previous section, regulatory bodies increasingly emphasize transparency for AI tools intended for clinical use. However, no common language has yet been established to describe or evaluate the quality of an explanation. Standardized evaluation frameworks, analogous to reporting guidelines such as Standards for Reporting Diagnostic Accuracy Studies,[59] are anticipated.

• **User-centric XAI design:** Current XAI methods are primarily designed to meet technical definitions of interpretability. A growing body of research advocates human-centered design, developing explanations that address clinicians' actual questions rather than optimizing mathematical properties. This requires close collaboration among AI developers, radiologists, and social scientists.

• **Integration into clinical workflows:** Integration of XAI into picture archiving and communication systems (PACS) and clinical decision support systems, with standardized presentation formats and user interface design, will be necessary for XAI to reach everyday clinical practice. This will pave the way for interactive explanations and human-in-the-loop systems, enabling users to query the model and receive tailored responses.

• **Beyond saliency—richer forms of explanation:** The field is moving beyond heatmaps toward richer explanation types, including textual reports, interactive what-if interfaces, and prototype-based comparisons. These more structured forms of explanation can provide clinicians with contextual, contrastive, and actionable information that supports genuine decision-making.

5.6 Concluding remarks

The question facing radiology is no longer whether AI will be integrated into clinical practice; it already has been. The real question is whether radiologists will serve as informed partners in AI-assisted diagnosis or as passive recipients of algorithmic outputs they cannot interrogate.

XAI does not fully solve the problems of clinical AI, and its limitations are real: post-hoc explanations of black-box models approximate the computational processes underlying their predictions, but they may not faithfully represent the internal features or interactions responsible for the output. There is an argument for preferring interpretable-by-design models whenever performance is comparable. Regardless of the explanation method used, trustworthy clinical AI depends on rigorous clinical and external validation, calibration, and ongoing human oversight.

This is not new for radiologists. Recognizing a false heatmap, questioning a vendor, identifying misplaced model confidence, and providing feedback to developers extend the appraisal skills the specialty already applies to every other diagnostic tool. XAI is what makes those skills actionable.

Footnotes

Conflict of interest disclosure

Gorkem Durak, Tugba Akinci D'Antonoli and Ulas Bagci serve as Section Editors for Diagnostic and Interventional Radiology. They had no involvement in the peer review of this article and had no access to information regarding its peer review. Halil Ertugrul Aktas declares no conflict of interest.

References

1. Durak G, Medetalibeyoglu A, Cicek V, Keles E, Bagci U. Beyond autonomy: why medicine needs artificial intelligence teammates, not artificial intelligence doctors. *Diagn Interv Radiol*. 2026. [Crossref]
2. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500-510. [Crossref]
3. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. [Crossref]
4. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *Radiographics*. 2017;37(7):2113-2131. [Crossref]
5. Jha D, Durak G, Sharma V, et al. A conceptual framework for applying ethical principles of AI to medical practice. *Bioengineering (Basel)*. 2025;12(2):180. [Crossref]
6. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115. [Crossref]
7. Kundu S. AI in medicine must be explainable. *Nat Med*. 2021;27(8):1328. [Crossref]
8. Dicle O, Atak F, Şenol AU, et al. Turkish Society of Radiology artificial intelligence applications guide: a roadmap to help navigate the artificial intelligence landscape. *Diagn Interv Radiol*. 2026;32(4):391-392. [Crossref]
9. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners - beyond saliency-based XAI approaches. *Eur J Radiol*. 2023;162:110786. [Crossref]
10. Haupt M, Maurer MH. Explainable artificial intelligence in radiology: methods, clinical applications, limitations, and future directions. *Eur J Radiol Artif Intell*. 2026;6:100098. [Crossref]
11. Ali S, Abuhmed T, El-Sappagh S, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf Fusion*. 2023;99:101805. [Crossref]
12. European Parliament, Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (EU Artificial Intelligence Act) [Internet]. 2024. [Crossref]
13. European Parliament, Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Off J Eur Union [Internet]. 2016;L119:1-88. [Crossref]
14. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15(11):e1002683. [Crossref]
15. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc ACM Conf Health Inference Learn (2020)*. 2020;2020:151-159. [Crossref]
16. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell*. 2021;3(7):610-619. [Crossref]
17. Hickman AJ, Gomes S, Warren LM, Smith NAS, Shenton-Taylor C. Assessing the generalisation of artificial intelligence across mammography manufacturers. *PLOS Digit Health*. 2025;4(8):e0000973. [Crossref]
18. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun*. 2019;10:1096. [Crossref]
19. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750. [Crossref]
20. Retzlaff CO, Angerschmid A, Saranti A, et al. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cogn Syst Res*. 2024;86:101243. [Crossref]

21. 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. [Crossref]
22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22-29; Venice, Italy. 2017. p. 618-626. [Crossref]
23. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); 2018 Mar 12-15; Lake Tahoe, NV, USA. IEEE; 2018. p. 839-847. [Crossref]
24. Bany Muhammad M, Yeasin M. Eigen-CAM: Visual explanations for deep convolutional neural networks. *SN Computer Science*. 2021;2(1):47. [Crossref]
25. Wang H, Wang Z, Du M, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14-19. IEEE; 2020. p. 111-119. [Crossref]
26. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*. 2015;10(7):e0130140. [Crossref]
27. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR. 2017;70:3319-3328. [Crossref]
28. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). New York (NY): Association for Computing Machinery; 2016. p. 1135-1144. doi:10.1145/2939672.2939778
29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4-9; Long Beach (CA). Red Hook (NY): Curran Associates Inc.; 2017. p. 4768-4777. doi:10.5555/3295222.3295230
30. Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Proceedings of the 35th International Conference on Machine Learning. PMLR; 2018;80:2668-2677. [Crossref]
31. Ghorbani A, Wexler J, Zou JY, Kim B. Towards automatic concept-based explanations. In: Advances in Neural Information Processing Systems. 2019;32:9273-9282. [Crossref]
32. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol. 2018;31(2):841-887. [Crossref]
33. Chen C, Li O, Tao C, Barnett AJ, Su J, Rudin C. This looks like that: deep learning for interpretable image recognition. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook (NY): Curran Associates Inc.; 2019. p. 8930-8941. [Crossref]
34. Aktas HE, Durak G, Bejar AM, et al. Uncertainty-aware explainable AI for pancreatic cysts: identifying deep learning vulnerabilities and ensuring safe clinical triage in IPMN management. *Res Sq [Preprint]*. 2026:rs.3.rs-9096790. [Crossref]
35. Faghani S, Moassefi M, Rouzrokh P, et al. Quantifying uncertainty in deep learning of radiologic images. *Radiology*. 2023;308(2):e222217. [Crossref]
36. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML); 2016. p. 1050-1059. [Crossref]
37. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). Red Hook (NY): Curran Associates Inc.; 2017. p. 6405-6416. [Crossref]
38. Guler TM, Ros PR, Erturk SM. Beyond pattern recognition: a Gödelian limit on self-validation in radiologic artificial intelligence. *J Am Coll Radiol*. 2026;23(6):1015-1016. [Crossref]
39. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS). Red Hook (NY): Curran Associates Inc.; 2018. p. 9525-9536. [Crossref]
40. Cerekci E, Alis D, Denizoglu N, et al. Quantitative evaluation of Saliency-Based Explainable artificial intelligence (XAI) methods in Deep Learning-Based mammogram analysis. *Eur J Radiol.* 2024;173:111356. [Crossref]
41. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. *Int J Comput Vis.* 2018;126(10):1084-1102. [Crossref]
42. Venkatesh K, Mutasa S, Moore F, Sulam J, Yi PH. Gradient-based saliency maps are not trustworthy visual explanations of automated AI musculoskeletal diagnoses. *J Imaging Inform Med*. 2024;37(5):2490-2499. [Crossref]
43. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*. 2023;307(4):e222176. [Crossref]
44. Fogliato R, Chappidi S, Lungren M, et al. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). New York (NY): Association for Computing Machinery; 2022. p. 1362-1374. [Crossref]
45. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. *J Am Coll Radiol*. 2010;7(9):698-704. [Crossref]
46. 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. [Crossref]
47. Wang AQ, Karaman BK, Kim H, et al. A framework for interpretability in machine learning for medical imaging. *IEEE Access*. 2024;12:53277-53292. [Crossref]
48. Petsiuk V, Das A, Saenko K. RISE: randomized input sampling for explanation of black-box models. arXiv [Preprint]. 2018;arXiv:1806.07421. [Crossref]
49. Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Netw Learn Syst*. 2017;28(11):2660-2673. [Crossref]
50. U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: guiding principles [Internet]. 2021 [cited 2025 Aug 30]. [Crossref]
51. U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Transparency for machine learning-enabled medical devices: guiding principles [Internet]. 2024 [cited 2025 Aug 30]. [Crossref]
52. Kotter E, Akinci D'Antonoli T, Cuocolo R, et al. Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. *Insights Imaging*. 2025;16(1):33. [Crossref]
53. Potočnik J, Fujs D. Navigating uncharted waters: select practical considerations in radiology AI compliance with the EU AI Act. *NPJ Digit Med*. 2025;8:630. [Crossref]
54. van Leeuwen KG, Doorn L, Gelderblom E. The AI Act: responsibilities and obligations for healthcare professionals and organizations. *Diagn Interv Radiol*. 2026;32(3):273-275.

[Crossref]

55. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. *arXiv [Preprint]*. 2023;arXiv:2303.08774. [Crossref]
56. Saab K, Tu T, Weng WH, et al. Capabilities of Gemini models in medicine. *arXiv [Preprint]*. 2024;arXiv:2404.18416. [Crossref]
57. Sharma S, Long J, Shih G, et al. CheXthought: a global multimodal dataset of clinical chain-of-thought reasoning and visual attention for chest X-ray interpretation. *arXiv [Preprint]*. 2026;arXiv:2604.26288. [Crossref]
58. Jha D, Durak G, Das A, et al. Ethical framework for responsible foundational models in medical imaging. *Front Med (Lausanne)*. 2025;12:1544501. [Crossref]
59. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 2015;277(3):826-832. [Crossref]