

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE: BUILDING TRUST IN AI-ASSISTED CLINICAL DECISION MAKING

Chatrapathi^[1], Punith J P^[2], J. P. Pramod^[3]

[1] B.Tech Student, Department of Computer Science and Engineering (CSE), VIT-Vellore, Tamil Nadu, India.
ORCID: 0009-0007-1950-8508

[2] B.Tech Student, Department of Computer Science and Engineering (CSE), NIT-Durgapur, India.
ORCID: 0009-0002-5925-709X

[3] Researcher, Department of R&D, The Foundry's Pvt. Ltd., Hyderabad, India.
ORCID: 0009-0001-1543-1394

ABSTRACT

AI is now commonplace in everyday healthcare: reading scans, predicting sepsis and alerting clinicians to the patients at greatest risk. By August 2024 the US FDA's public list of AI-enabled medical devices had reached 950 entries. However, the most accurate AI models tend to be the most difficult to interpret. Some are so complex that even their builders cannot explain how the model arrived at a particular conclusion. This is the "black box" problem, and it matters in medicine, where an inexplicable AI decision can lead to negative health consequences for patients. This article looks at explainable AI (XAI), the suite of tools and techniques that try to open up that black box so that doctors, patients and regulators can understand and trust what AI systems are telling them. We cover the spectrum of techniques, from interpretable, human-readable models to post-hoc tools such as SHAP, LIME and heat-maps; explain why explainability is critical in a hospital; examine what regulatory data show about gaps in transparency; and provide a framework for building trust among AI, doctors and patients. We conclude that explainability cannot be a one-time checklist item before a product goes to market, but must be a core pillar of how AI is monitored and used in healthcare.

KEYWORDS: AI Regulation, Clinical Decision Support, Clinician Trust, Explainable AI, Interpretability, LIME, Medical Imaging, Predictive Diagnostics, SHAP

1. INTRODUCTION

Walk into almost any modern hospital and AI is quietly at work somewhere in the building. Radiologists get a second opinion from an algorithm before they finish reading a scan. Sepsis-prediction tools run in the background of electronic health record (EHR) systems, watching for early warning signs. Emergency rooms use AI to help decide which patients need attention first. None of this was common even five years ago. The U.S. Food and Drug Administration (FDA) list of AI-enabled medical devices grew from 692 entries in late 2023 to 882 by March 2024 and 950 by August 2024 [10] (Table 1 and Figure 3).

Computer assistance in medical diagnostics is not new. Rule-based scoring methods have long supported clinicians; cardiovascular risk scores have existed since the 1960s, giving an assessment from a small number of weighted factors such as smoking, cholesterol and blood pressure. In the 1970s, MYCIN, an advisory system for antibiotic therapy, generated explicit if-then rules. When such a system makes a decision, the doctor can see what was decided and why, which keeps it trustworthy even when it errs.

A modern deep learning system is different. No fixed set of rules is written by human engineers. Instead, the algorithm builds its own internal model by finding statistical patterns in massive training datasets. This lets it discover nuances in patient histories and images that engineers could never have written into explicit rules, but it is also exactly why its logic is so hard to interpret. No concise sentence adequately describes what such a model does.

Despite this, these systems are effective enough to have been authorized by the FDA, adopted in clinics worldwide and used on real patients today. Yet approval is not trust. A treating physician must decide whether to act on a recommendation or cross-examine it; a hospital administrator is concerned with liability and whether a doctor can explain an AI-driven decision that turns out to be wrong; a patient may struggle to accept a machine influencing decisions about their health; and a regulator must decide, before any patient is involved, that the evidence behind a model's claims justifies the risk.

These difficulties are what drive the field of explainable AI (XAI). When a model predicts a patient's risk of sepsis, the accuracy of the prediction matters, but so does the doctor's ability to see what drove it. Without that, an alarm only tells the physician "this number is high", which means little without a way to check it against the actual patient.

It is worth clarifying what "trust" means here. The goal is not to maximize clinicians' trust in an AI system but to *calibrate* it, so that trust is proportionate to the system's actual performance. Too little trust wastes the assistance AI can offer; too much trust (automation bias) is dangerous, because physicians may accept a bad recommendation simply because it looks sophisticated. Explainability is what allows calibration, by letting the doctor compare the clinical information at hand with the model's forecast.

The rest of this article is organized as follows. Section 2 covers the trade-off between performance and transparency and the main families of explanation methods. Section 3 explains why explanation matters in practice. Sections 4 and 5 give case studies in two major fields: medical imaging and prediction from health records. Section 6 examines what the published regulatory data indicate about transparency. Section 7 outlines a framework for building trust between humans and medical AI, and Section 8 covers the main barriers to adoption and ways to address them. To ground the discussion in open data, Table 1 shows the pace of growth in FDA-listed AI/ML devices.

Table 1. Cumulative FDA-listed AI/ML-enabled medical devices at selected checkpoints.

Checkpoint	Cumulative AI/ML devices	Source
Late 2023	692	FDA list, Oct 2023 update
Mar 2024	882	FDA list, May 2024 update (devices through Mar 2024)
Aug 2024	950	FDA list update, Aug 2024 (as reported by MDR Regulator)

2. THE BLACK-BOX PROBLEM AND A TAXONOMY OF EXPLAINABILITY

Here is the core problem: the AI models that perform best in healthcare, deep neural networks trained on millions of images or patient records, are also the hardest to understand. They find patterns across thousands or millions of internal parameters, and no human could realistically inspect all of that and explain it in plain terms. This creates a trade-off: the models that do best on

complex clinical prediction tasks tend to explain themselves worst, while simpler, more transparent models usually give up some accuracy in exchange. Figure 1 shows this trade-off conceptually.

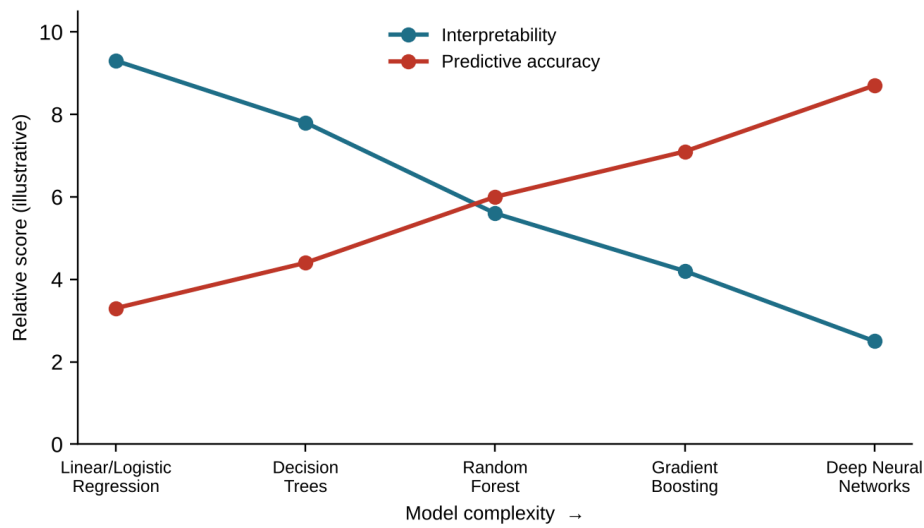


Figure 1. The accuracy–interpretability trade-off across common classes of clinical AI models. (Original conceptual illustration, not plotted from measured data; it shows the general pattern described in the reviewed literature, not exact values for any specific model.)

Explainability approaches generally fall into two broad camps, shown in Figure 2. The first is *intrinsic interpretability*: models that are transparent by design, like logistic regression or decision trees, where you can trace how an input leads to an output. The second is *post-hoc explainability*: techniques applied after a complex model has been trained, which try to shed light on its reasoning without changing the model itself.

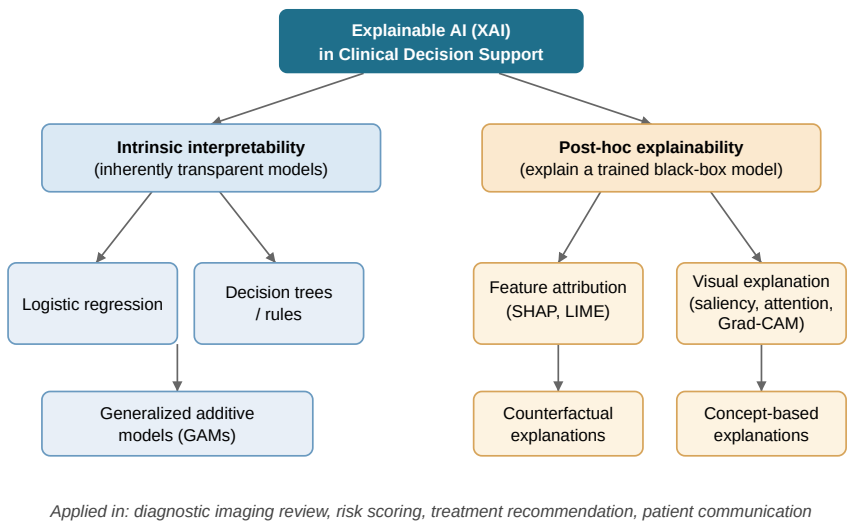


Figure 2. A taxonomy of explainable AI approaches used in clinical decision support systems. (Original conceptual diagram based on the technique categories in the cited literature, not derived from a specific dataset.)

In most industries, this trade-off is just an inconvenience. In medicine, it is a real dilemma. If a doctor cannot understand why an algorithm recommended withholding a treatment, or flagged a scan as malignant, they are stuck: override a system that might be right, or defer to a system whose reasoning they cannot see. Liability, patient safety and professional judgment all collide at that moment.

3. WHY EXPLAINABILITY MATTERS IN CLINICAL PRACTICE

It is easy to lump explainability under "ethics" or "compliance" and move on. Yet it offers clear, practical benefits in everyday care:

- **Detecting errors:** Explanations provide clues that a model is right for the wrong reason, such as a dermatology classifier that associated rulers or ink markings in clinical photographs with melanoma rather than the lesion itself.
- **Calibrating trust:** Transparency moderates the clinician's faith in a recommendation. According to a 2024 systematic review, explanations generally increased trust where they were simple, succinct and clinically meaningful; elsewhere they had no effect, or even reduced it. They can also create automation bias when they look scientific but are empty noise [1]. Trust has also been identified as a major psychological variable in whether clinicians take an AI recommendation to heart in practice [2].
- **Accountability and documentation:** When a doctor's diagnosis or recommended treatment is challenged, a convincing rationale is crucial for medical records, peer review and liability defense.
- **Regulation:** Medical AI is broadly treated as "high risk" under the EU AI Act and thus falls under requirements for transparency, human oversight and record-keeping. The Total Product

Life Cycle approach that the FDA is developing for AI/ML-based Software as a Medical Device similarly pushes toward continuous evidence on how a model behaves long after initial clearance.

4. EXPLAINABILITY IN MEDICAL IMAGING AND DIAGNOSTICS

Medical imaging remains the largest application domain for clinical AI, and it is where explanation techniques are most mature. Trained on large annotated datasets, convolutional neural networks (CNNs) can now perform pulmonary nodule detection on CT, mammographic classification and diabetic retinopathy diagnosis, with sensitivity and specificity often meeting or exceeding human readers on narrow, well-defined tasks. Reviews of XAI in medical image analysis find saliency and Grad-CAM-style methods dominant [4]. A broad systematic review across the medical and healthcare domain also highlighted problems with standardization of evaluation and integration into practice [3].

The two dominant forms of explanation for images are saliency maps and attention-based visualizations (such as Grad-CAM), which highlight the pixels the model prioritized for classification. A radiologist can use them to check whether the model is focused on a detected tumor or on an incidental artefact such as a scanner label or padding marker. However, saliency maps are approximations, not proofs of causal reasoning, and critics have shown that visually similar saliency maps can persist even when the model's weights have been altered, which contributes to skepticism about using visual explanation in isolation. A 2024 systematic review of medical image analysis documented this and other limitations, noting that many visual explanation techniques may not meet requirements for accuracy and clinical usability at the same time [4].

5. EXPLAINABILITY IN PREDICTIVE DIAGNOSTICS AND EHR-BASED RISK MODELS

Predictive models built on EHR data, whether predicting sepsis onset, hospitalization or deterioration, use tabular, longitudinal data rather than images, so feature-attribution methods are a natural fit. SHAP (SHapley Additive exPlanations), based on game theory, assigns each input feature a value for its contribution to a particular prediction. In a sepsis risk model, SHAP may show a clinician that an alert was driven mainly by the patient's current heart rate and lactate values rather than age and medical history, which the clinician can readily check against the patient. LIME (Local Interpretable Model-agnostic Explanations) does a similar job but explains one prediction at a time: it fits a simple, interpretable model to approximate the complex model's behavior in the neighborhood of that single prediction.

Counterfactual explanations, a third approach, map directly onto how clinicians talk about thresholds: for example, "a 20% lower creatinine level would have prevented the model from flagging high acute kidney injury risk." They mirror clinical discussion of decision thresholds for intervening, making them appealing for shared decision-making. A 2024 scoping review of EHR-based predictive models found feature-attribution methods, particularly SHAP, to be the most prevalent, followed by rule- and example-based systems [5]. An earlier scoping review of EHR-based models covering 2009–2019 charted similar trends and noted that explanation quality was rarely evaluated, as opposed to merely provided [6].

A perennial limitation of SHAP, counterfactual explanations and feature-attribution tools generally is that models capture correlations, not the true causal mechanisms of medicine. This has given rise to concern about "explainability theater": an explanation can provide a deceptive facade of understanding, a risk more insidious than openly acknowledging that a model is opaque. A systematic review of XAI properties and of how well clinicians use explanations likewise found limited quantitative evaluation of how well doctors comprehend and use them, and few formal evaluations of explanation quality in clinical settings [7].

6. THE REGULATORY AND EVIDENTIARY LANDSCAPE

Regulatory authorization of AI-enabled medical devices has taken off. Figure 3 tracks the cumulative count on the FDA list at the checkpoints in Table 1: 692 devices in late 2023, 882 through March 2024 and 950 by August 2024 [10].

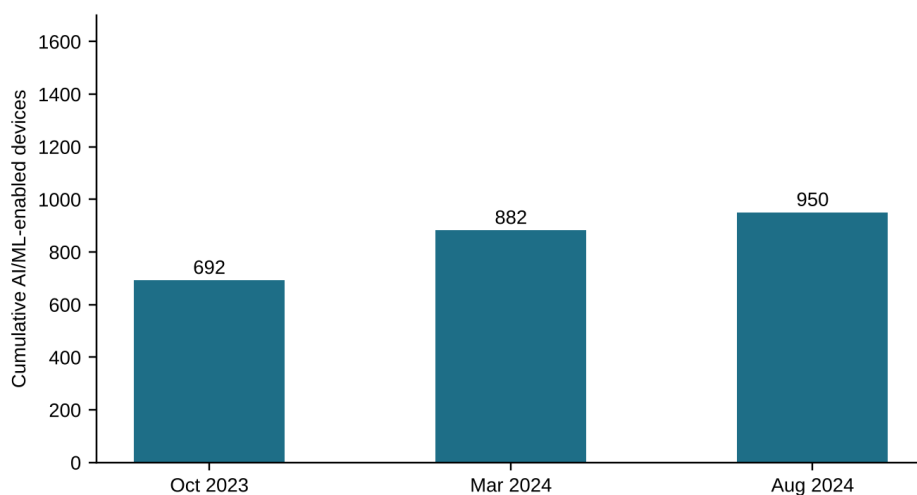


Figure 3. Cumulative FDA-listed AI/ML-enabled medical devices at the checkpoints in Table 1, late 2023 to August 2024. Source: FDA list updates [10].

Authorizations have grown faster than transparency reporting has kept up. Public device summaries frequently omit subgroup performance and full validation data, exactly the information a doctor needs to judge whether to trust a result for a specific patient. Predetermined Change Control Plans (PCCPs), which let manufacturers pre-specify how a model may update itself after reaching the market, are only beginning to appear, though adoption is expected to rise as the pathway matures.

Table 2 summarizes the main regulatory and evidence gaps facing explainable clinical AI.

Table 2. Regulatory and evidence gaps in explainable clinical AI.

Challenge	Description	Clinical / regulatory impact
Data privacy & security	AI systems process highly sensitive patient information across imaging, genomic and EHR sources.	Breach risk undermines patient trust and creates legal exposure.
Algorithmic bias	Models trained on non-representative datasets can underperform for under-represented populations [8].	Unequal or unsafe outcomes across demographic groups.
Interpretability ("black box")	Deep learning decision processes are not directly inspectable.	Clinicians hesitate to adopt; accountability becomes unclear.
Regulatory fragmentation	Divergent requirements across the FDA, EU AI Act and other national frameworks.	Slows cross-border deployment and standardization.
Transparency reporting gaps	Many device summaries omit performance by subgroup or full validation data.	Limits clinicians' ability to judge when to trust a given output.

7. BUILDING TRUST: A PRACTICAL FRAMEWORK

Across both theory and practice, a set of principles guides development toward appropriate, sustainable trust between AI, clinicians and patients, shown as a loop in Figure 4. Reviews of the literature show that technical explainability and practical trust-building are distinct, though partly intertwined, objectives: explanation methods are widely provided, but their quality and clinical usefulness are rarely evaluated [6, 7].

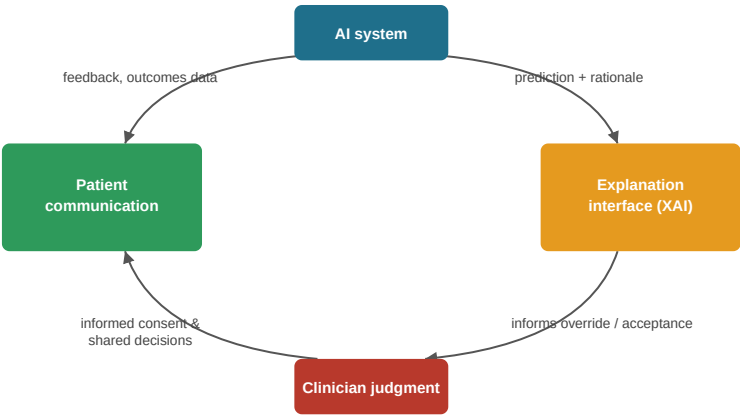


Figure 4. The clinical trust loop: how explainability connects AI output to clinician judgment and patient communication. (Original conceptual framework.)

The principles are as follows:

- **Match the explanation to the user:** a radiologist, a hospital administrator and a patient each need a different form and depth of explanation.

- **Show uncertainty, not just output:** confidence intervals and out-of-distribution warnings do more for trust calibration than a single confident-looking number.
- **Validate explanations, not just predictions:** an explanation method should be tested for faithfulness to the model's actual reasoning, not merely for plausibility.
- **Keep a human meaningfully in the loop:** explainability is most valuable when paired with workflows that genuinely allow override, rather than ones where disagreement is procedurally discouraged.
- **Treat explainability as continuous, not a one-time certification:** as models are updated post-deployment under frameworks such as PCCPs, their explanations require ongoing re-validation.

8. CHALLENGES AND RECOMMENDATIONS FOR ADOPTION

Despite substantial technical progress, several structural challenges continue to limit the safe and equitable adoption of explainable AI in clinical settings. Table 3 summarizes practical measures aligned to each challenge area.

Table 3. Recommended measures for adopting explainable clinical AI.

Measure	Purpose	Expected outcome
Federated learning	Train models across institutions without centralizing raw patient data.	Protects privacy; expands training diversity without data-transfer risk.
Explainable AI (XAI) requirements	Require faithful, validated explanations prior to clinical deployment.	Builds clinician and patient trust; supports accountability.
Standardized transparency reporting	Mandate subgroup performance and validation disclosure in device summaries.	Enables informed clinical judgment about when to trust a given output.
Human-in-the-loop workflow design	Structure interfaces so override remains genuinely accessible, not discouraged.	Preserves clinical judgment; reduces automation bias.
Continuous post-market validation	Re-validate model and explanation performance after deployment updates.	Maintains safety and interpretability across the model's lifecycle.

9. CONCLUSION

Explainable AI remains unsolved in the clinical space and may never be "solved" in the sense that every deep learning model is perfectly transparent. That is not the real objective. The objective is calibrated trust: giving clinicians, patients and regulators enough understanding of a model's reasoning to know when it can be trusted and when to reject its suggestion outright.

As AI moves from identification toward more consequential clinical decisions (from suggesting that a scan looks like cancer to suggesting which treatments could alleviate the disease), insight, rather than accuracy alone, will determine its real utility to patient care. This is particularly well supported in precision medicine, where a 2024 systematic review of explainable AI in digital health repeatedly identified transparency and valid explanations as essential to turning individualized, data-driven

treatment decisions into clinically feasible strategies [9]. Consistent progress requires standard mechanisms for transparency reporting, explanation methods that are validated rather than merely plausible, and regulation that treats explainability as an evolving property of the system in its environment, rather than a set-and-forget approval.

REFERENCE MAPPING

Table 4 maps each reference to the theme or section it supports, so the connection between the literature reviewed and the arguments made is traceable.

Table 4. Reference mapping.

Ref.	Theme / section supported
[1]	Clinician trust: how explanations increase or decrease trust (Section 3)
[2]	Clinician-focused trust in healthcare AI (Section 3)
[3]	Broad XAI landscape across medical and healthcare domains (Section 4)
[4]	XAI in medical image analysis (Section 4)
[5]	XAI applications in EHR research, 2024 scoping review (Section 5)
[6]	XAI applied to real-world EHR data, 2009–2019 (Section 5)
[7]	Explanation properties, quality and effectiveness evaluation (Sections 5, 7)
[8]	Fairness and bias in healthcare AI (Table 2)
[9]	XAI in precision medicine and digital health (Conclusion)
[10]	FDA list of AI-enabled medical devices (Introduction, Section 6)

REFERENCES

1. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI*. 2024;3:e53207. doi:10.2196/53207.

2. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res*. 2020;22(6):e15154. doi:10.2196/15154.

3. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The Enlightening Role of Explainable Artificial Intelligence in Medical & Healthcare Domains: A Systematic Literature Review. *Comput Biol Med*. 2023;166:107555. doi:10.1016/j.combiomed.2023.107555.

4. Muhammad D, Bendechange M. Unveiling the Black Box: A Systematic Review of Explainable Artificial Intelligence in Medical Image Analysis. *Comput Struct Biotechnol J*. 2024;24:542–560. doi:10.1016/j.csbj.2024.08.005.

5. Caterson J, Lewin A, Williamson E. The Application of Explainable Artificial Intelligence (XAI) in Electronic Health Record Research: A Scoping Review. *Digit Health*. 2024;10. doi:10.1177/20552076241272657.

6. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z. Explainable Artificial Intelligence Models Using Real-World Electronic Health Record Data: A Systematic Scoping Review. *J Am Med Inform Assoc*. 2020;27(7):1173–1185. doi:10.1093/jamia/ocaa053.

7. Jung J, Lee H, Jung H, Kim H. Essential Properties and Explanation Effectiveness of Explainable Artificial Intelligence in Healthcare: A Systematic Review. *Heliyon*. 2023;9(5):e16110. doi:10.1016/j.heliyon.2023.e16110.
8. Ueda D, Kakinuma T, Fujita S, et al. Fairness of Artificial Intelligence in Healthcare: Review and Recommendations. *Jpn J Radiol*. 2024;42(1):3–15. doi:10.1007/s11604-023-01478-z.
9. Allen B. The Promise of Explainable AI in Digital Health for Precision Medicine: A Systematic Review. *J Pers Med*. 2024;14(3):277. doi:10.3390/jpm14030277.
10. US Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices (list). Updates of Oct 2023 and Aug 2024. fda.gov.