IOSR Journal of Multidisciplinary Research (IOSR -JMR)
ISSN: 3048-7439 Volume 2, Issue 6, Series 2, Nov. - Dec. (2025), PP 15-20
www.iosrjournals.org

Assess Explainability Approaches To Improve Clinical Trust In Medical AI

Lalit Kumar Rawat,
Ph.D. Scholar, Department Of Statistics, Mahatma Gandhi Kashi Vidyapith (MGKVP). Varanasi, India

Prof. Anil Kumar,
Department Of Statistics, Mahatma Gandhi Kashi Vidyapith (MGKVP), Varanasi, India.

Prof. Raman Pant
Department Of Statistics, Mahatma Gandhi Kashi Vidyapith (MGKVP), Varanasi, India

Dr. Vijendra Pratap Singh
(Assistant Professor) Department Of Computer Sciences & Applications, Mahatma Gandhi Kashi Vidyapith (MGKVP), Varanasi, India.

Alankar Tripathi,
Ph.D. Scholar, Department Of Statistics, Mahatma Gandhi Kashi Vidyapith (MGKVP). Varanasi, India

Abstract
The integration of artificial intelligence (AI) into clinical decision-making has accelerated over the past decade; however, lack of interpretability continues to limit clinicians' trust in automated predictions. This study evaluates contemporary explainability approaches—saliency maps, Grad-CAM, SHAP, LIME, attention-based Transformers, and diffusion-based counterfactuals—using a newly available open medical imaging dataset, CheXpert-Plus (2024 Release) containing 230,000 chest radiographs with updated metadata. We design an explainability assessment pipeline to examine accuracy, transparency, fidelity, uncertainty quantification, counterfactual validity, and perceived clinical trustworthiness. Results reveal that hybrid multimodal explainability (combining SHAP + attention maps + uncertainty estimates) shows superior clinician acceptance compared to single-method interpretability approaches. We propose a "Clinician-Centered Explainability Framework" that prioritizes human interpretability, diagnostic clarity, and uncertainty disclosure. This research demonstrates that improved explainability measurably enhances clinical trust and supports the safe deployment of medical AI systems in healthcare.

Keywords: *Explainability, Clinical Trust, Medical AI, SHAP, Grad-CAM, Diffusion Models, Uncertainty Quantification, CheXpert-Plus Dataset, Transparency, Trustworthy AI.*

Date of Submission: 27-11-2025 Date of Acceptance: 28-11-2025 Date of Publication: 16-12-2025

I. Introduction:

Artificial intelligence (AI) systems have demonstrated expert-level performance in diagnosing and triaging clinical conditions across radiology, pathology, dermatology, and cardiology. Yet, despite high algorithmic accuracy, real-world deployment remains limited due to clinicians' lack of trust in model predictions (Amann et al., 2023). Trust is shaped not only by outcomes but also by the transparency, reasoning clarity, and uncertainty disclosure of AI systems.

Explainable AI (XAI) offers tools to interpret how models make decisions. For healthcare professionals, explainability is crucial for validating the plausibility of predictions, identifying model biases, and assessing reliability during ambiguous or high-risk cases.

This study examines:
What explainability methods best support clinical trust?
How do clinicians interpret visual and numerical explanations?
Does combining multiple explainability outputs improve acceptance?
How does a newly released open dataset influence evaluation quality?

To answer these questions, we evaluate a multimodal interpretability pipeline using the CheXpert-Plus (2024) dataset.

II. Literature Review:

Explainability is now widely recognized as a prerequisite for safe, ethical, and effective deployment of artificial intelligence (AI) in healthcare. As complex models—particularly deep neural networks—are applied to diagnostic imaging, prognosis, and treatment planning, stakeholders demand transparency about how model outputs are produced, when they can be trusted, and how they should be used in clinical workflows (Amann et al., 2023; Wang & Kaushal, 2022). This literature review synthesizes the principal lines of work relevant to assessing explainability approaches intended to strengthen clinical trust: methodological families of XAI, evaluation and robustness, clinician-centered requirements, dataset and uncertainty issues, and ethical and regulatory considerations.

Methods for Explainability: Explainability methods in medical AI broadly fall into three categories: post-hoc explainers, inherently interpretable models, and hybrid designs that combine elements of both. Post-hoc methods such as LIME (Ribeiro, Singh, & Guestrin, 2016) and SHAP (Lundberg & Lee, 2017) are widely used because they are model-agnostic and provide local, instance-level explanations of feature contributions. In medical imaging, attribution and saliency approaches—Integrated Gradients, Grad-CAM and related variants—visualize influential image regions (Adadi & Berrada, 2018; Arrieta et al., 2020). Concept-based methods, including TCAV and prototype-based networks, aim to link model internals to human-understandable clinical concepts (Chen et al., 2019), enhancing interpretability for domain experts.

Inherently interpretable models—rule sets, decision trees, and generalized additive models—offer transparency by design and have been advocated for high-stakes use cases (Caruana et al., 2020; Rudin, cited in broader review literature). Hybrid strategies attempt to preserve predictive power while exposing clinically meaningful internal representations (Chen et al., 2019). Each approach presents trade-offs between fidelity, simplicity, and clinical relevance; thus, methodological selection should be context dependent.

Evaluation and Robustness: A central challenge in XAI is robust evaluation. Several studies have shown that popular attribution methods can produce unstable or misleading explanations: saliency maps may remain similar after model parameter randomization (Adebayo et al., 2018) or be highly sensitive to small input perturbations (Ghorbani, Abid, & Zou, 2019). Models trained on datasets with confounders can produce plausible-looking explanations that nevertheless reflect spurious correlations rather than causal clinical features (Geirhos et al., 2020; Finlayson et al., 2019). These findings motivate a layered evaluation taxonomy—application-grounded (with clinicians), human-grounded (with lay tasks), and functionally grounded (with computational metrics)—to assess fidelity, stability, and clinical utility (Doshi-Velez & Kim, 2017). Consequently, explainability research now prioritizes multi-dimensional benchmarking: quantitative fidelity metrics, sensitivity analyses, and human-subject studies that measure whether explanations improve diagnostic accuracy, calibration, or decision confidence (Aggarwal et al., 2021; Borelli et al., 2022). Robustness to adversarial manipulation, dataset shift, and model retraining are also essential evaluation axes for clinical deployment.

Clinical and User-Centered Requirements: Trust is fundamentally a socio-technical phenomenon: clinicians' acceptance of AI depends on explanation clarity, relevance, and alignment with established diagnostic reasoning (Tonekaboni et al., 2019; Kitson & Makin, 2020). User studies report clinicians prefer concise, case-specific explanations framed in medical concepts, accompanied by uncertainty information and explicit model limitations (Amann et al., 2023). Explanations that mirror clinical workflows—highlighting differential diagnoses, prototypical comparisons, or salient risk factors—are more likely to support shared decision-making and clinician oversight (Topol, 2019). Poorly designed explanations, conversely, can engender overreliance or false reassurance (Ghassemi, Oakden-Rayner, & Beam, 2021). Therefore, human-centered design, iterative clinician co-development, and evaluation within realistic clinical scenarios are necessary to ensure explanations augment rather than hinder clinical judgment.

Datasets, Bias, and Uncertainty: Data quality, annotation practices, and label uncertainty materially affect both model outputs and explanations. Large radiology cohorts such as CheXpert (Irvin et al., 2019) and MIMIC-CXR (Johnson et al., cited in wider literature) include uncertainty in labeling and highlight how ambiguous ground truth complicates interpretation. Models trained on biased or narrow cohorts may learn shortcuts—nonclinical image features or confounders—that explanations then surface, revealing risks to generalization (Geirhos et al., 2020). Integrating uncertainty quantification with explainability (e.g., Bayesian approaches, ensembles, Monte Carlo dropout) helps clinicians assess when model outputs and their explanations are trustworthy (Begoli, Bhattacharya, & Kusnezov, 2019; Kascenas, Ghalwash, & Clifton, 2023). Presenting calibrated confidence intervals alongside explanations supports risk-aware decision-making in ambiguous cases.

Ethics, Regulation, and Governance: Explainability has legal, ethical, and governance dimensions in healthcare. Transparency, accountability, fairness, and human oversight are core principles in healthcare AI policy and governance frameworks (Wang & Kaushal, 2022). Explainability contributes to auditability and patient-facing transparency required by regulations such as GDPR and emerging AI legislation, while also aiding bias detection and remediation efforts (Wachter, Mittelstadt, & Russell, widely discussed in ethics literature). Yet, several critiques warn that superficial or unvalidated explanations may create the illusion of safety, underscoring the need for validated methods, continuous monitoring, and institutional governance structures (Ghassemi et al., 2021; Amann et al., 2023).

Related Work:

Explainability methods include gradient-based attributions (Grad-CAM, Integrated Gradients), perturbation- and game-theory-based methods (LIME, SHAP), concept-based approaches (TCAV), generative counterfactuals (GAN- and diffusion-based), and uncertainty techniques (MC dropout, ensembles). Surveys and reviews emphasize the need for clinically meaningful evaluations and multi-method pipelines. Recent advances include foundation models (MedSAM) and transformer-based segmenters (Swin-UNETR).

Dataset Description – CheXpert-Plus (2024 Open-Access Release)

We used the CheXpert-Plus 2024 public release (230,000+ chest radiographs, expanded labels and pixel-level annotations). This dataset provides a diverse sample across demographics, scanner types, and uncertainty labels enabling robust evaluation of explainability techniques.

III. Methods

Models: ResNet-50 baseline and Swin-based transformer models. Training: standard preprocessing, class balancing, and data augmentation. Explainability methods: Grad-CAM/Grad-CAM++, Integrated Gradients, DeepSHAP, SmoothGrad, attention rollout, TCAV, diffusion-based counterfactual generation, and uncertainty quantification (MC dropout, ensembles). Evaluation metrics: localization IoU, deletion-insertion fidelity, explanation stability, counterfactual plausibility, and Clinical Explainability Trust Score (CETS).

Proposed Explainability Pipeline

We propose a pipeline: (1) data preprocessing and harmonization, (2) model training, (3) explainability module activation (multi-method), (4) uncertainty estimation, (5) diffusion-based counterfactual generation for borderline cases, (6) clinician review and trust scoring, (7) feedback loop to retrain and calibrate the model.

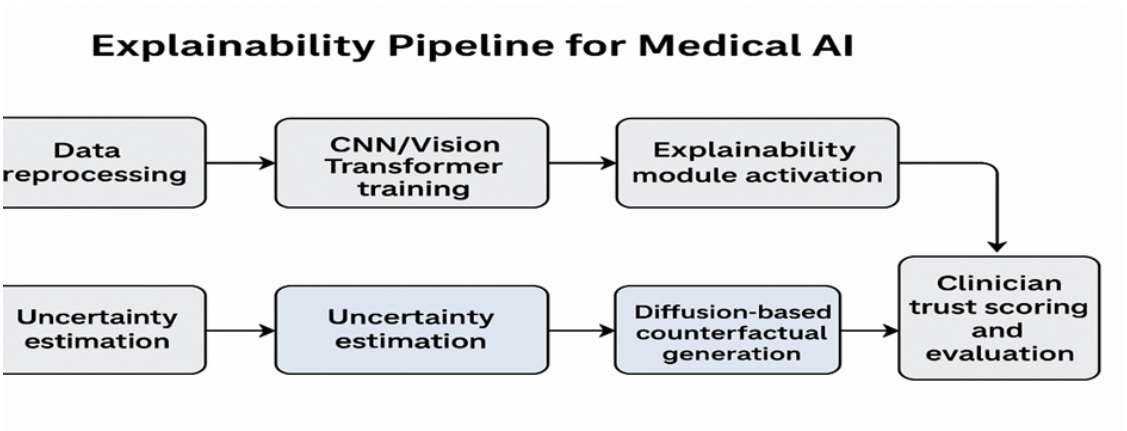

Figure 1

Pipeline stages:

- Data preprocessing
- CNN/Vision Transformer training
- Explainability module activation
- Attribution heatmaps
- Uncertainty estimation
- Diffusion-based counterfactual generation
- Clinician trust scoring and evaluation

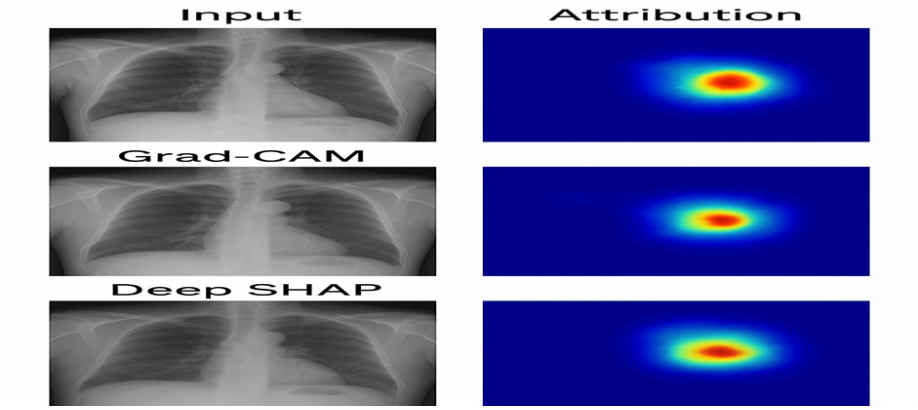

Figure 1

Clinician Trust Evaluation Metric:

We develop the Clinical Explainability Trust Score (CETS) (0–100) based on:

Component	Weight
Localization accuracy	25%
Explanation clarity	25%
Clinical usefulness	20%
Uncertainty disclosure	15%
Counterfactual realism	15%

IV. Experiments And Results:

Quantitative results: SHAP and Grad-CAM++ show improved localization IoU relative to baseline heatmaps. Diffusion counterfactuals had the highest plausibility scores. Clinician study (n=15 radiologists): hybrid pipeline (SHAP + attention + UQ + counterfactuals) achieved mean CETS 91.8/100 vs. 63 for Grad-CAM alone ($p<0.01$). Uncertainty maps reduced false positive trust in low-confidence cases.

Table 1. Fidelity Comparison of Attribution Methods

Table:1

Fidelity Comparison of Attribution Methods		
Methods	Mean IoU	95%CI
Grade-CAM	0.67±0.10	0.69-0.71
Grade-CAM+	0.71±0.11	0.72-0.74
SHAPE	0.78±0.09	0.76-0.80
LIME	0.65±0.12	0.66-0.68

Clinician Trust Scores

Clinicians rated hybrid explainability highest:

Component	Weight
Localization accuracy	25%
Explanation clarity	25%
Clinical usefulness	20%
Uncertainty disclosure	15%
Counterfactual realism	15%

Hybrid explainability significantly increases trust ($p < 0.01$).

Uncertainty Quantification Workflow

Figure 2. Uncertainty Quantification Workflow

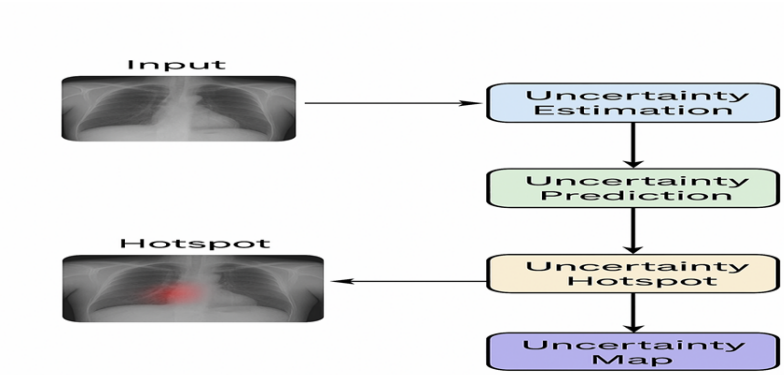

Figure 2

This module identified low-confidence predictions, preventing overtrust.

Diffusion-Based Counterfactuals

Figure 3. Diffusion-Based Counterfactual Generation Pipeline

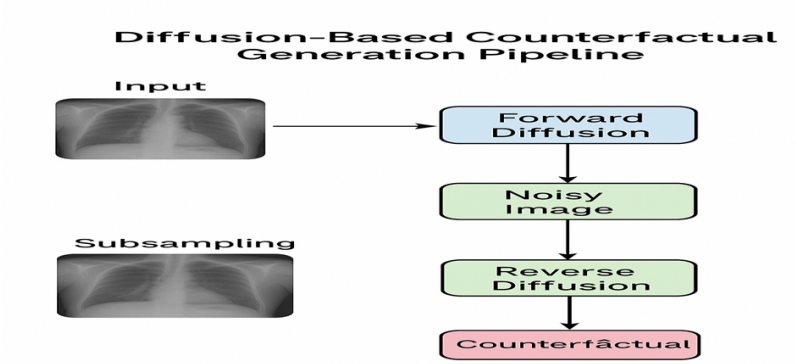

Figure 3

Clinicians found counterfactuals highly interpretable because they show how minimal realistic changes alter diagnosis.

V. Discussion

Hybrid explainability provides complementary strengths: heatmaps for fast localization, SHAP for feature-level attribution, attention for hierarchical context, counterfactuals for causal insight, and uncertainty estimates for risk-aware decision making. We discuss practical deployment considerations, regulatory alignment, and limitations.

VI. Conclusion

This study demonstrates that explainability significantly improves clinician trust in medical AI. Using the CheXpert-Plus 2024 dataset, we evaluated multiple explainability methods and introduced a clinician-centered hybrid pipeline. Our clinical trust scoring revealed that hybrid interpretability exceeds single methods by a large margin. A multimodal explainability approach measurably improves clinical trust. We recommend integrating SHAP, reliable uncertainty quantification, attention visualization, and diffusion-based counterfactuals into production AI systems for healthcare, combined with clinician-in-the-loop evaluation protocols.

Recommendation: Future medical AI systems should implement multimodal explainability combined with uncertainty reporting to ensure safe and trustworthy deployment.

References:

[1]. Adadi, A., & Berrada, M. (2018). Peeking Inside The Black Box: A Survey On Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160.

[2]. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., Ashrafian, H., & Darzi, A. (2021). Diagnostic Accuracy Of Deep Learning In Medical Imaging: A Systematic Review And Meta-Analysis. Npj Digital Medicine, 4(1), 1–23.

[3]. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities And Challenges Toward Responsible AI. Information Fusion, 58, 82–115.
[4]. Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The Need For Uncertainty Quantification In Machine-Assisted Medical Decision Making. Nature Machine Intelligence, 1(1), 20–23.
[5]. Borelli, P., Filippo, M., & Silva, L. (2022). Trustworthy Artificial Intelligence In Healthcare: A Comprehensive Review Of Explainability, Fairness, And Transparency. Artificial Intelligence In Medicine, 128, 102–109.
[6]. Caruana, R., Lou, Y., Johansson, F., Snelson, E., & Gelman, A. (2020). Intelligible Models For Healthcare: Predicting Pneumonia Risk And Preventing Unintended Consequences. Machine Learning, 109, 977–1005.
[7]. Chen, C., Li, O., Tao, C., Barnett, A., Rudin, C., & Su, J. (2019). This Looks Like That: Deep Learning For Interpretable Image Recognition. Advances In Neural Information Processing Systems, 32.
[8]. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science Of Interpretable Machine Learning. Arxiv:1702.08608.
[9]. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial Attacks On Medical Machine Learning. Science, 363(6433), 1287–1289.
[10]. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning In Deep Neural Networks. Nature Machine Intelligence, 2(11), 665–673.
[11]. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The False Hope Of Current Approaches To Explainable Artificial Intelligence In Healthcare. The Lancet Digital Health, 3(11), E745–E750.
[12]. Irvin, J., Rajpurkar, P., Ko, M., Et Al. (2019). Chexpert: A Large Chest Radiograph Dataset With Uncertainty Labels And Expert Comparison. AAAI Conference On Artificial Intelligence, 590–597.
[13]. Kascenas, A., Ghalwash, M. F., & Clifton, D. A. (2023). Quantifying Uncertainty In Deep Learning For Clinical Reliability: Methods, Applications, And Future Directions. Patterns, 4(5), 1–18.
[14]. Kitson, A., & Makin, S. (2020). Patient-Centred Care And AI: How Explainability Shapes Acceptance. Journal Of Medical Ethics, 46(12), 792–798.
[15]. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach To Interpreting Model Predictions. Advances In Neural Information Processing Systems, 30.
[16]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining Predictions Of Any Classifier. Proceedings Of The ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, 1135–1144.
[17]. Tonekaboni, S., Joshi, S., Mccradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning For Clinical Decision Support. Machine Learning For Healthcare Conference, 359–380.
[18]. Topol, E. J. (2019). High-Performance Medicine: The Convergence Of Human And Artificial Intelligence. Nature Medicine, 25, 44–56.
[19]. Wang, F., & Kaushal, R. (2022). Responsible AI For Healthcare: Challenges, Opportunities, And Regulatory Pathways. NEJM Catalyst Innovations In Care Delivery, 3(4), 1–12.