



Explainable Artificial Intelligence in Healthcare: Methods, Benefits, and Challenges for Trustworthy Clinical Decision Support

Author: **Unnati Mehta**

Program: M.Sc. Artificial Intelligence

College: Kandivali Education Society's B.K Shroff college of Arts and M.H Shroff college of commerce(Autonomous)

unnatimehta97@gmail.com

How to Cite this Article:

Mehta, U. (2026). Explainable Artificial Intelligence in Healthcare: Methods, Benefits, and Challenges for Trustworthy Clinical Decision Support. International Journal of Creative and Open Research in Engineering and Management, 2(10), 1-9.
<https://doi.org/10.55041/ijcope.v2i10.065>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i10.065>

Abstract—Artificial intelligence (AI) systems are increasingly studied for tasks such as medical-image analysis, risk prediction, and clinical decision support. However, strong predictive performance does not by itself show why a model reached a particular conclusion or whether that conclusion is appropriate for a patient. This paper reviews explainable artificial intelligence (XAI) methods that can make healthcare predictions more understandable to clinicians and other stakeholders. A focused narrative review was conducted using foundational work on local explanations and Shapley-value-based feature attribution, a healthcare-focused review, and World Health Organization guidance on AI ethics. The paper compares model-specific and model-agnostic explanations, with particular attention to LIME, SHAP, and visual explanation methods. The review indicates that explanations can support error analysis, communication, and model auditing, but they should not be treated as proof that a prediction is correct, fair, or clinically safe. Explanation quality depends on data quality, model behaviour, the intended user, and the clinical context. The paper proposes a practical evaluation framework combining predictive performance, explanation fidelity, stability, clinical usefulness, fairness checks, and human oversight. It concludes that XAI should be integrated into the full lifecycle of clinical AI systems rather than added only after model development.

Index Terms—Explainable AI, healthcare, clinical decision support, interpretability, SHAP, LIME.

I. INTRODUCTION

Healthcare decisions often combine clinical measurements, patient history, imaging, laboratory results, and professional judgement. Machine-learning models can detect patterns in these data and may assist with screening, diagnosis, prognosis, and prioritisation. Yet many high-performing models, especially complex ensemble and deep-learning systems, are difficult to interpret directly. A prediction such as “high risk” may be useful as a warning, but it is less useful if a clinician cannot determine which evidence influenced the result or whether the model relied on an inappropriate signal.

Explainable artificial intelligence (XAI) describes methods intended to make model behaviour or individual predictions more understandable. In healthcare, this is not merely a presentation issue. Explanations may help users investigate errors, compare a prediction with clinical knowledge, communicate uncertainty, and identify possible reliance on biased or irrelevant features. The World Health Organization (WHO) stresses transparency, accountability, human autonomy, and safety in the design and use of AI for health [1].

This paper addresses the following question: How can current XAI methods improve the interpretability of AI-assisted healthcare decisions, and what limitations must be considered before they are used in clinical practice? Its contribution is a focused comparison of common explanation approaches and a practical evaluation framework for assessing them in context. The paper does not claim to validate a clinical model or report new patient-level experiments.

II. REVIEW METHOD AND SCOPE

A focused narrative literature review was used. The source set was selected to cover three complementary areas: foundational explanation techniques, research reviewing XAI in medical applications, and ethical guidance for health AI. The main sources include Ribeiro et al.'s work on Local Interpretable Model-agnostic Explanations (LIME) [2], Lundberg and Lee's SHapley Additive exPlanations (SHAP) framework [3], a review of interpretability

methods in medical health applications [4], a broader systematic review of XAI for healthcare [5], and WHO guidance [1].

The review compares methods by the kind of explanation they provide, their dependence on a particular model, their potential use in clinical workflows, and their limitations. It is not a systematic review: no claim is made that every relevant publication was retrieved, and the sources were not assessed through a formal risk-of-bias protocol. The conclusions should therefore be read as a structured synthesis of selected literature rather than a quantitative estimate of clinical effectiveness.

III. WHAT EXPLAINABILITY MEANS IN HEALTHCARE

Interpretability and explainability are related but not identical concepts. An interpretable model, such as a small decision tree or a carefully specified linear model, may be understandable from its structure. An explanation method, by contrast, attempts to describe the behaviour of a model that may itself be difficult to inspect. In everyday use, the terms are sometimes used interchangeably, so a paper or system should state what type of understanding it intends to provide.

Explanations can be global or local. A global explanation summarises patterns across a model—for example, which variables tend to influence predictions overall. A local explanation focuses on one output, such as why a model assigned a particular patient a higher estimated risk. These answers are not interchangeable: a feature that is influential on average may not be important for an individual case. Likewise, an explanation of a model's behaviour is not automatically a causal explanation of a disease.

Healthcare also involves multiple audiences. A data scientist may need technical details about feature attribution, a clinician may need a concise account of relevant evidence and uncertainty, and a patient may need plain-language information about how a result should be interpreted. A technically detailed explanation can still fail if it is not understandable to the person expected to use it.



IV. COMMON XAI METHODS

A. Local Interpretable Model-agnostic Explanations (LIME)

LIME explains an individual prediction by creating perturbed examples around the case of interest and fitting a simpler, interpretable model that approximates the original model locally [2]. In a clinical example, the explanation might indicate that a set of measured variables contributed to a model's elevated risk estimate. Its main advantage is flexibility: the original predictor does not need to be a particular model family.

The local nature of LIME is also a limitation. Results can depend on how nearby examples are generated, how features are represented, and which interpretable model is fitted. If small changes in sampling produce different explanations, a clinician may receive an unstable account of the same prediction. LIME output should therefore be checked for stability and should not be presented as a definitive medical rationale.

B. SHapley Additive exPlanations (SHAP)

SHAP assigns feature-importance values using ideas from Shapley values in cooperative game theory. The method expresses a prediction as a baseline plus contributions attributed to individual features [3]. Depending on the model and implementation, SHAP can support local explanations and summaries across many cases. This can help analysts inspect whether a model repeatedly relies on clinically plausible information or appears to use suspicious proxies.

SHAP values are contributions to a model output under a defined explanation setup; they do not establish that a feature causes the outcome. Correlated variables, the choice of background data, and assumptions about feature dependence can affect interpretation. Computation can also be demanding for some models. Consequently, SHAP plots should be accompanied by a description of the output being explained and the assumptions used to calculate contributions.

C. Visual Explanations for Medical Images

For image-based models, methods such as gradient-weighted class activation mapping (Grad-CAM) highlight regions of an image associated with a model's output. In radiology, a heat map may help a reviewer inspect whether a model attends to a plausible anatomical region. Visual explanations can be intuitive, but a heat map is not a diagnosis and does not prove that the model has learned the correct clinical concept. A model may focus on image borders, labels, acquisition artefacts, or other shortcuts.

D. Intrinsically Interpretable Models

A further option is to choose a model whose logic can be inspected directly, such as a sparse linear model, a small decision tree, or a rule-based system. These models may be easier to audit and explain. However, simple structure does not guarantee good calibration, fairness, or clinical validity, and some tasks may require more complex representations. Model selection should consider the clinical purpose rather than assuming that either maximum complexity or maximum simplicity is always best.

V. APPLICATIONS IN CLINICAL WORKFLOWS

In medical imaging, explanations can direct attention to regions that influenced a classification and support the review of false positives and false negatives. In risk prediction, feature contributions may help a clinician understand why a patient was flagged for closer follow-up. In triage and decision-support systems, a concise explanation may help users decide when to seek additional evidence or override a recommendation. These are potential uses, not evidence that an explanation by itself improves patient outcomes.

XAI may also support model development and governance. Global summaries can reveal unexpected dependence on variables that encode care access or local workflow rather than underlying health. Case-level explanations can help investigate model failures. When performance changes after deployment, explanation patterns may provide clues about data drift. However, such clues require follow-up analysis; explanation tools do not replace conventional validation, monitoring, or clinical review.

VI. CHALLENGES AND RISKS

A. Faithfulness and Stability

An explanation is useful only to the extent that it reflects the behaviour of the model it describes. A plausible-looking explanation may be misleading if it changes substantially under minor perturbations or does not accurately approximate the model. Evaluations should therefore test fidelity to the predictor and stability across reasonable changes in input, sampling, and implementation.

B. Accuracy Is Not the Same as Clinical Safety

A model can achieve strong aggregate performance while failing on underrepresented patient groups or in a different hospital. Similarly, a readable explanation does not correct poor data quality, label errors, distribution shift, or an unsuitable target. Predictive metrics, calibration, subgroup performance, and external validation remain necessary. Explanations are one component of assurance, not a substitute for it.

C. Bias, Privacy, and Accountability

Clinical datasets may reflect unequal access to care, historical treatment patterns, and differences in measurement. If a model learns these patterns, an explanation may expose a concern but cannot resolve it on its own. Patient data must also be protected when explanations are generated, stored, or shared. WHO guidance emphasises human autonomy, transparency, accountability, inclusiveness, and the protection of individuals affected by health AI [1]. Responsibility for clinical decisions should remain clear, including the role of clinicians, institutions, developers, and regulators.

D. Human Factors and Overtrust

An explanation can increase confidence without increasing correctness. Users may over-trust a persuasive visual or treat a list of important features as proof that a recommendation is justified. Explanations should be tested with intended users, use language appropriate to their expertise, and communicate uncertainty.



Training should make clear that an explanation describes a model's behaviour rather than guaranteeing the truth of its output.

Any such study would need careful data governance and should avoid presenting model outputs as medical advice.

VII. A PRACTICAL EVALUATION FRAMEWORK

Based on the reviewed literature, this paper proposes that healthcare XAI systems be evaluated across six dimensions. First, predictive performance should be measured on data representative of the intended setting. Second, explanation fidelity should assess whether the explanation reflects the model's behaviour. Third, stability should test whether reasonable input or sampling changes cause unjustified shifts in the explanation. Fourth, clinical usefulness should be assessed with intended users through structured tasks rather than assumed from attractive visualisations. Fifth, fairness and privacy checks should examine performance across relevant groups and the handling of sensitive data. Sixth, operational governance should define oversight, documentation, escalation, and monitoring after deployment.

These dimensions should be considered together. For example, an explanation that is stable but clinically unhelpful may not improve a workflow, while a highly understandable explanation that misrepresents model behaviour may be actively harmful. The evaluation plan should be set before deployment and tailored to the specific clinical use case, intended users, and consequences of error.

Dimension	Example check	Why it matters
Predictive validity	External test data; calibration	Checks whether outputs are dependable.
Faithfulness	Compare explanation with model response	Checks whether explanation reflects the predictor.
Stability	Repeat under small, valid perturbations	Finds fragile explanations.
Clinical usefulness	User study with clinicians	Tests whether explanations aid decisions.
Fairness and privacy	Subgroup analysis; access controls	Addresses unequal harms and data exposure.
Governance	Audit logs and post-deployment review	Supports accountability over time.

VIII. DISCUSSION

The literature suggests that no single XAI method is suitable for every healthcare task. LIME offers local, model-agnostic approximations; SHAP provides a principled feature-attribution framework; visual methods can support image inspection; and intrinsically interpretable models may make the underlying logic easier to audit. Their outputs answer different questions, and each method has assumptions that must be made visible.

A central practical lesson is that explainability should be designed around a decision, not around a plot. Before choosing a method, developers should specify who needs the explanation, what decision it should support, what level of detail is appropriate, and how a user can challenge the output. A clinician may need a concise explanation with uncertainty and relevant patient context, while a model auditor may need global feature analysis and subgroup checks. The same explanation should not automatically be used for both.

The present review has limitations. It is a focused narrative synthesis rather than a reproducible systematic review, and it does not compare methods on a shared clinical dataset. It therefore cannot establish which technique performs best or quantify the effect of explanations on patient outcomes. Future work could compare LIME, SHAP, and visual explanations on a publicly available, appropriately licensed dataset, using pre-defined metrics for fidelity, stability, subgroup performance, and clinician usability.

IX. CONCLUSION

Explainable AI can help make healthcare models easier to inspect, question, and govern, but explanation should not be confused with proof of correctness. LIME, SHAP, visual attribution methods, and interpretable models each provide different forms of insight and carry distinct limitations. In clinical settings, explanations should be evaluated for fidelity, stability, usability, fairness, privacy, and operational accountability alongside predictive performance and external validation. The most responsible approach is to treat XAI as part of a broader human-centred safety process, with clinicians retaining appropriate oversight and with clear procedures for monitoring and correcting system failures.

REFERENCES

- [1] World Health Organization, "Ethics and governance of artificial intelligence for health: WHO guidance," Geneva, Switzerland: WHO, 2021. [Online]. Available: <https://www.who.int/publications/i/item/9789240029200>
- [2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774. [Online]. Available: <https://arxiv.org/abs/1705.07874>
- [4] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, and H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," Informatics in Medicine Unlocked, vol. 40, Art. no. 101286, 2023, doi: 10.1016/j.imu.2023.101286.
- [5] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," Computer Methods and Programs in Biomedicine, vol. 226, Art. no. 107161, 2022, doi: 10.1016/j.cmpb.2022.107161.