Implementing Explainable AI (XAI) in Diagnostic Workflows

By: Mihran Ahmed Khan

Email: mydulkhan@gmail.com

Abstract: As of 2026, Artificial Intelligence has transitioned from a theoretical novelty to a fundamental utility in the global medical sector. However, a significant barrier remains: the "black box" nature of complex algorithms, which often provide life-critical diagnoses without interpretable reasoning [6]. This lack of transparency has led to a documented "trust gap," where a majority of clinicians hesitate to implement AI-driven recommendations that contradict their professional intuition [1]. This research proposal explores the transition from pure predictive accuracy to Explainability i.e ability to explain the diagnosis results. The goal is to develop a framework for Explainable AI (XAI) that provides clinicians with visual and textual justifications for AI outputs. By prioritizing human-centered design and mitigating algorithmic bias [7], this study seeks to enhance diagnostic safety and ensure that AI integration is both ethical and equitable across diverse global health systems.

Introduction: The integration of AI in the medical sector is no longer a future prospect but a present reality. By early 2026, AI adoption across healthcare organizations has expanded rapidly, ranging from automated radiology screenings to real-time predictive analytics for patient deterioration [4]. Despite this rapid adoption, the "implementation gap" persists: models that perform exceptionally in controlled lab settings often fail when faced with the "domain shift" of real-world hospitals [2]. This research, authored by Mihran Ahmed Khan, proposes a fundamental shift in medical AI development. Rather than focusing solely on increasing raw accuracy, this study investigates how "Explainability" features such as heatmaps, feature importance charts, and confidence scores can bridge the gap between machine logic and clinical expertise [4]. By moving towards a "Human-in-the-Loop" model, the research seeks to transform AI from an opaque diagnostic tool into a transparent partner that assists physicians in making more informed, safe, and accountable decisions [1].

Research Question: How can the addition of explainable features such as visual heatmaps, feature importance breakdowns, and confidence scores help clinicians better trust, understand, and act on AI-generated diagnoses, and does this vary depending on the clinical setting or patient population being served?

Literature Review: Recent advancements in Machine Learning (ML) have achieved diagnostic parity with specialists in high-stakes fields like oncology and cardiology [4]. However, the literature increasingly highlights the "trust deficit" in clinical settings, where physicians demand to know the "why" behind a machine's decision before taking action [1][2]. Recent studies underscore the necessity of “Federated Learning” to incorporate diverse global data without compromising patient privacy, yet these models often remain uninterpretable to the end user [3]. Furthermore, the growing deployment of AI in clinical workflows makes the need for a standardized "Explainability Protocol" more urgent than ever to prevent undetected algorithmic errors [4][6].

Methodology:

- **Framework Design:** Development of a "Transparency Interface" that utilizes SHAP aka SHapley Additive exPlanations [5] to highlight specific features such as "nodule density" or "tissue irregularity" that justify an AI-generated diagnosis.
- **Dataset Harmonization:** Partnering with diverse healthcare networks to test model resilience against "domain shift" using multi-ethnic and multi-regional data [2][3].
- **Human-in-the-Loop (HITL) Simulation:** A controlled experiment where clinicians are presented with AI results with and without explainability features to measure the impact on diagnostic speed and error reduction [1][7].

Conclusion: At its core, this research is about closing a gap that has quietly grown alongside the rise of medical AI — the gap between what a machine concludes and what a clinician can actually understand and trust. The tools already exists to make a more transparent diagnosis ecosystem; what has been missing is a structured, human-centered framework for bringing those tools into real clinical workflows in a way that is fair, testable, and scalable. That is what this study sets out to build. Starting with controlled simulations and expanding across diverse patient populations, the goal is not simply to make AI more explainable in theory, but to prove that explainability makes a measurable difference in how doctors diagnose and how safely they do it. If this research succeeds, it will not just improve one diagnostic tool — it will offer a model for how the entire healthcare industry can begin to move away from blind algorithmic trust and toward a more accountable, human-centered partnership with AI.

References/Works Cited:

[4] Abbas, Qaiser, et al. "Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges." *Healthcare*, vol. 13, no. 17, 29 Aug. 2025, pp. 2154–2154, www.mdpi.com/2227-9032/13/17/2154, https://doi.org/10.3390/healthcare13172154.

[3] Abbas, Syed Raza, et al. "Federated Learning in Smart Healthcare: A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration." *Healthcare*, vol. 12, no. 24, 22 Dec. 2024, pp. 2587–2587, www.mdpi.com/2227-9032/12/24/2587, https://doi.org/10.3390/healthcare12242587.

[2] Anjara, Sabrina G, et al. "Examining Explainable Clinical Decision Support Systems with Think Aloud Protocols." *PloS One*, vol. 18, no. 9, 14 Sept. 2023, pp. e0291443–e0291443, https://doi.org/10.1371/journal.pone.0291443.

[5] Lundberg, Scott, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *ArXiv:1705.07874 [Cs, Stat]*, 24 Nov. 2017, arxiv.org/abs/1705.07874.

[7] Rezaeian, Olya, et al. "The Impact of AI Explanations on Clinicians Trust and Diagnostic Accuracy in Breast Cancer." *ArXiv.org*, 2024, arxiv.org/abs/2412.11298.

[1] Rosenbacke, Rikard, et al. "How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review." *JMIR AI*, vol. 3, 30 Oct. 2024, pp. e53207–e53207, ai.jmir.org/2024/1/e53207, https://doi.org/10.2196/53207.

[6] Xu, Hanhui, and Kyle Michael. "Medical Artificial Intellegence and the Black Box Problem – a View Based on the Ethical Principle of "Do No Harm."" *Intelligent Medicine*, vol. 4, no. 1, 1 Aug. 2023, www.sciencedirect.com/science/article/pii/S2667102623000578, https://doi.org/10.1016/j.imed.2023.08.001.