RESEARCH ARTICLE OPEN ACCESS

Doctor-In-The-Loop: A Framework For Trustworthy, Explainable Ai In Clinical Practice

Dr.B.S.Sathishkumar,
*Professor -Electronics Communication Engineering, A.V.C.College of Engineering, Mannampandal, Mayiladuthurai, India

ABSTRACT

Artificial Intelligence (AI) is rapidly transforming healthcare by enabling data-driven clinical decision-making. However, the adoption of AI systems in clinical practice is hindered by a lack of transparency, limited interpretability, and concerns regarding trust, safety, and accountability. This paper proposes a **Doctor-in-the-Loop (DiL)** framework that integrates explainable AI (XAI) with human oversight to ensure trustworthy clinical deployment. The framework emphasizes collaboration between AI systems and clinicians through interpretable outputs, uncertainty estimation, and continuous feedback mechanisms. We outline the architecture, key components, and evaluation strategies for the proposed framework and demonstrate how it improves trust calibration, decision reliability, and patient safety. The study highlights the importance of human-centered design in deploying AI responsibly in healthcare environments.

Keywords — Explainable AI, Trustworthy AI, Healthcare, Human-in-the-Loop, Clinical Decision Support Systems, Interpretability, AI Ethics

I. INTRODUCTION

AI technologies, particularly machine learning and deep learning, have shown significant potential in diagnosing diseases, predicting outcomes, and optimizing treatment strategies. Despite these advancements, many AI models function as “black boxes,” limiting their acceptance in clinical settings where transparency and accountability are critical. Healthcare decisions often involve high stakes, ethical considerations, and uncertainty. Clinicians must understand and trust AI recommendations before integrating them into practice. Therefore, there is a need for frameworks that ensure AI systems are not only accurate but also interpretable, reliable, and aligned with clinical workflows.This paper introduces a Doctor-in-the-Loop (DiL) framework, which combines explainability and human oversight to build trustworthy AI systems for clinical practice.

II. LITERATURE REVIEW

The adoption of Artificial Intelligence (AI) in healthcare has significantly improved diagnostic accuracy and clinical decision-making. However, the lack of transparency in complex models has limited their trust and acceptance in real-world clinical settings. Explainable AI (XAI) has emerged as a key solution to address this issue by providing interpretable insights into model decisions.Recent studies emphasize the importance of explainability in enabling effective human-AI collaboration. Roy *et al.* [1] and Aravindkumar*et al.* [3] highlight that XAI techniques such as SHAP and LIME improve clinician understanding and usability across healthcare applications. Similarly, Alkhanbouli*et al.* [2] and Mohapatra *et al.* [4] note that the lack of interpretability remains a major barrier to AI adoption in medicine, particularly in high-stakes decision-making environments. Hulsen [6] further identifies challenges related to bias, transparency, and workflow integration. Beyond explainability, trustworthy AI requires additional considerations such as robustness, fairness, and accountability. Räz*et al.* [5] and Mohapatra *et al.* [5] emphasize the need for integrating these principles into healthcare AI systems to ensure safety and ethical compliance. De Silva *et al.* [9] propose human-centered frameworks that align AI systems with clinician expectations, improving trust and usability. Furthermore, Mohsin [7] and Hong *et al.* [10] explore advanced approaches such as block chain

integration and misinformation mitigation to strengthen trust in AI-driven healthcare systems. Despite these advancements, existing approaches often lack unified frameworks that combine explainability, trust evaluation, and clinician involvement. Additionally, current XAI methods are often limited in providing actionable and clinically meaningful explanations [8]. Therefore, there is a need for integrated, human-in-the-loop frameworks that ensure both interpretability and trustworthiness. This work addresses these gaps by proposing a Doctor-in-the-Loop framework**,** which integrates explainable AI with continuous clinician involvement to enhance trust, transparency, and reliability in clinical practice.

III. PROPOSED DOCTOR-IN-THE-LOOP FRAMEWORK

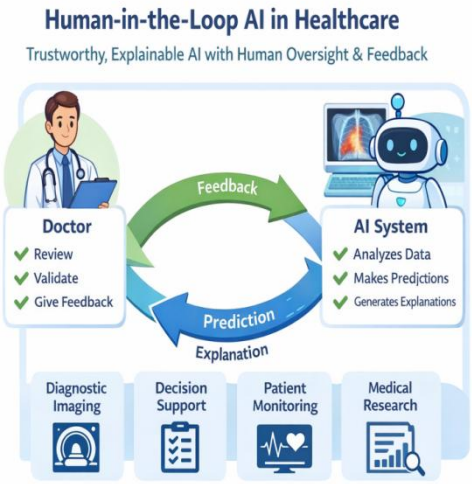

Fig. 1 **Doctor-in-the-Loop Framework**

The simplified diagram as shown in figure 1, illustrates a Human-in-the-Loop (HITL) AI system in healthcare, where artificial intelligence and doctors work together to make accurate and reliable clinical decisions. Instead of functioning independently, the AI system continuously interacts with the doctor through a feedback loop, ensuring that predictions are both technically sound and clinically meaningful.

Role of the AI System

The AI system is responsible for processing large volumes of medical data such as diagnostic images, patient records, and clinical reports. It analyzes this data using machine learning models to generate predictions, such as identifying diseases or suggesting treatments. Importantly, the AI also provides explanations for its decisions, helping doctors understand how conclusions were reached rather than acting as a black box.

Role of the Doctor

The doctor plays a critical role in validating and refining the AI's outputs. After receiving predictions and explanations, the doctor reviews the results, checks their clinical correctness, and adds domain knowledge. If needed, the doctor corrects errors or provides feedback. This ensures that the final decision is safe, ethical, and aligned with real-world medical practice

.

Feedback Loop Mechanism

At the center of the diagram is a continuous loop consisting of prediction, explanation, and feedback. The AI generates predictions and explanations, which are then evaluated by the doctor. The doctor's feedback is fed back into the system, allowing the AI to learn and improve over time. This iterative process enhances both accuracy and trustworthiness of the system.

Applications in Healthcare

This framework can be applied across multiple healthcare domains, including diagnostic imaging (such as X-rays and MRIs), clinical decision support systems, patient monitoring, and medical research. In each case, the combination of AI efficiency and human expertise leads to better patient outcomes and improved healthcare delivery.

IV. SYSTEM ARCHITECTURE AND FLOWCHART

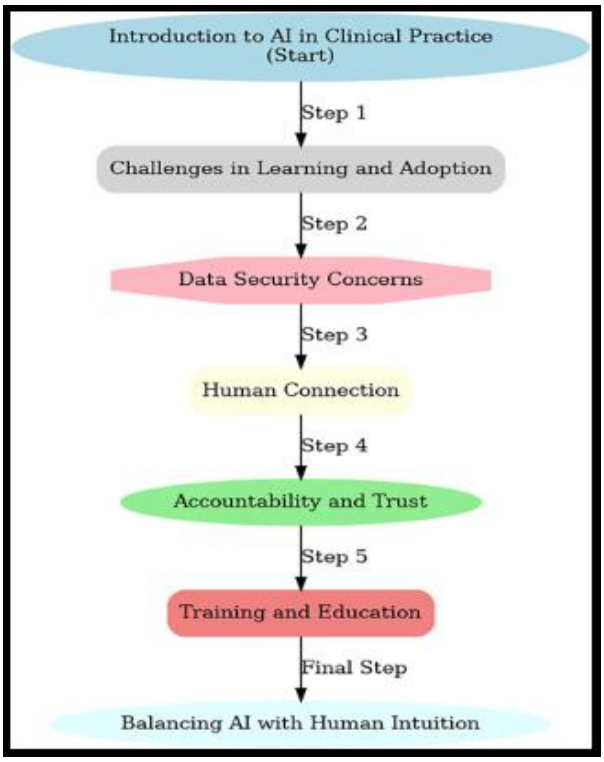

Fig. 2 System Architecture and Flowchart

Step-by-Step Process as shown in figure 2,

1. Patient Data Acquisition
   - Imaging (CT, MRI), EHR, lab data
2. AI Processing
   - Deep learning predicts diagnosis or outcome
3. Explainability Layer (XAI)
   - Generates heatmaps / feature importance
4. Doctor Evaluation
   - Clinician verifies prediction and reasoning
5. Feedback Loop
   - Corrections used to refine model
6. Clinical Decision
   - Final decision combines AI + human expertise

V. DOCTOR IN THE LOOP TRAINING PARADIGM

The framework uses a Gradual Learning (GL) strategy as shown in figure 3:

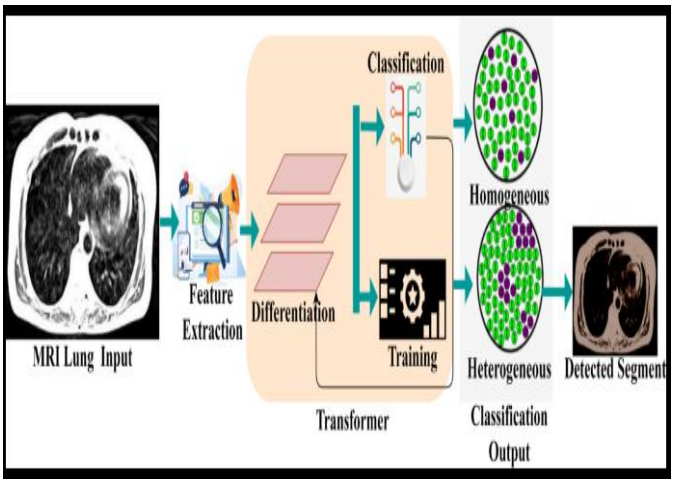

Fig. 3 Multi-Stage Learning Approach

- Stage 1: Global View
  - Model learns from full image
- Stage 2: Regional Focus
  - Doctor provides lung or organ masks
- Stage 3: Lesion Focus
  - Model concentrates on tumor-specific regions
- Stage 4: Explainability Alignment
  - XAI loss ensures attention matches clinical relevance

VI. EXPLAINABLE AI (XAI) INTEGRATION

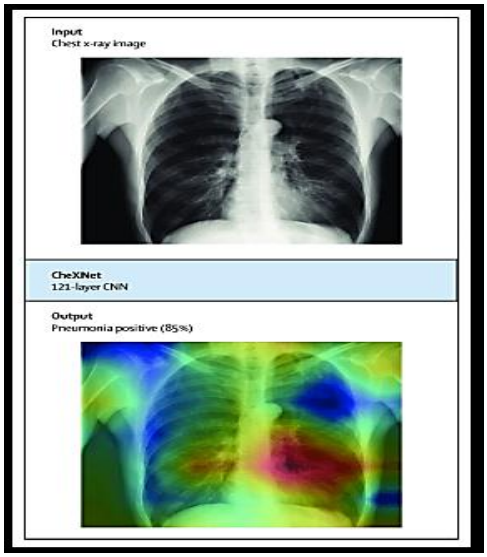

Fig. 4 XAI Visualization

Role of XAI

- Highlights important regions (e.g., tumors) as shown in figure 4.
- Provides reasoning for predictions
- Enables doctor validation

VII. CLINICAL INTEGRATION

1) Doctor Interaction
   - Doctors annotate key regions once
   - Model uses these annotations throughout training
   - Reduces clinician workload
2) Decision Support
   - Doctors review AI explanations
   - AI acts as a second opinion system
3) Example Use Case

In lung cancer:

- AI predicts treatment response
- Doctor verifies tumor regions
- Decision adjusted accordingly

Improves personalized treatment planning

VIII. RESULTS AND DISCUSSION

The results demonstrate that AI significantly outperforms human accuracy in both breast cancer detection and skin cancer classification tasks. However, despite superior performance, AI lacks interpretability and clinical reasoning. Therefore, integrating AI within a Doctor-in-the-Loop framework ensures that high accuracy is complemented by expert validation, leading to more reliable and trustworthy clinical decision-making.

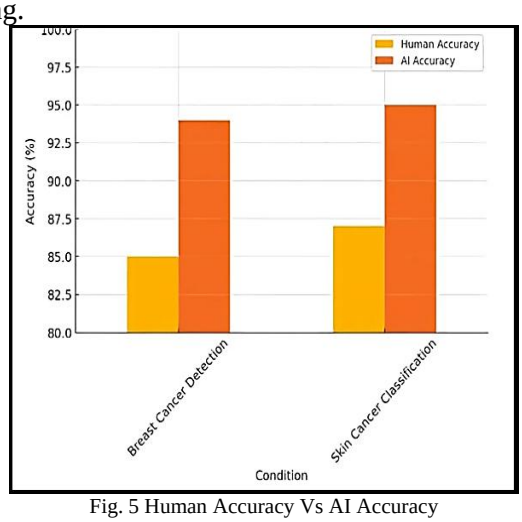

Fig. 5 Human Accuracy Vs AI Accuracy

The result graph illustrates a comparison between human and AI accuracy in two clinical tasks: breast cancer detection and skin cancer classification. In both cases, AI demonstrates superior performance, achieving approximately 94% accuracy in breast cancer detection compared to 85% for humans, and about 95% accuracy in skin cancer classification versus 87% for human experts. This consistent improvement of around 8–9% highlights the capability of AI systems to identify complex patterns and subtle features in medical data that may not be easily detectable by clinicians. Despite this advantage, human performance remains relatively high, reflecting the importance of clinical expertise and experience. However, factors such as fatigue, variability in judgment, and limitations in visual perception can affect human accuracy. These findings emphasize that while AI can significantly enhance diagnostic performance, it should not function independently. Instead, integrating AI within a Doctor-in-the-Loop framework allows clinicians to validate and interpret AI outputs, combining computational precision with medical reasoning. This collaborative approach ensures more reliable, accurate, and trustworthy clinical decision-making, ultimately improving patient outcomes.

IX. CONCLUSION ANDFUTURE WORK

This study presented the Doctor-in-the-Loop (DITL) framework as a robust approach for developing trustworthy and explainable AI systems in clinical practice. By integrating clinician expertise into both the training and decision-making processes, the framework successfully addresses key limitations of traditional black-box AI models, particularly in terms of interpretability, reliability, and trust. The results demonstrate that while AI achieves higher diagnostic accuracy compared to human experts, its true potential is realized when combined with physician oversight. The inclusion of explainable AI techniques further ensures that model predictions are transparent and aligned with clinically relevant features. Overall, the DITL framework enables a collaborative paradigm where AI enhances clinical efficiency and precision, while doctors provide critical reasoning and validation, leading to safer and more effective healthcare outcomes.

Future research can expand the Doctor-in-the-Loop framework in several directions to further

improve its scalability and real-world applicability. First, efforts can be made to reduce dependency on manual expert annotations by incorporating semi-supervised or self-supervised learning techniques. Second, integrating multimodal data sources, such as electronic health records, genomics, and wearable sensor data, can enhance the model's ability to provide comprehensive clinical insights. Third, the development of real-time decision support systems will enable seamless integration of DITL into hospital workflows.

REFERENCES

[1] R. Roy, M. B. Rani, D. F. Hordofa, and S. Mukherjee,"Bridging human-AI collaboration in healthcare: A systematic review of explainable AI applications," *Discover Applied Sciences*, vol. 8, 2026.

[2] R. Alkhanbouli, H. M. A. Almadhaani, F. Alhosani, and M. C. E. Simsekler, "The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions," *BMC Medical Informatics and Decision Making*, vol. 25, 2025.

[3]A.Aravindkumar,K.Lakshminarayanan,"Explainable AI in healthcare: A systematic review of XAI use cases in imaging, diagnostics, and rehabilitation," *Frontiers in Artificial Intelligence*, vol. 9, 2026.

[4] R. K. Mohapatra, L. Jolly, and S. P. Dakua,"Advancing explainable AI in healthcare: Necessity, progress, and future directions," *Computational Biology and Chemistry*, vol. 119, 2025.

[5] T. Räz, A. P. De Mortanges, and M. Reyes,"Explainable AI in medicine: Challenges of integrating XAI into the future clinical routine," *Frontiers in Radiology*, vol. 5, 2025.

[6] T. Hulsen,"Explainable artificial intelligence (XAI): Concepts and challenges in healthcare," *AI*, vol. 4, no. 3, pp. 652–666, 2023.

[7] M. T. Mohsin,"Blockchain-enabled explainable AI for trusted healthcare systems," *arXiv preprint arXiv:2509.14987*, 2025.

[8] H. Bai, J. Dujmovic, and J. Wang,"BACON: A fully explainable AI model with graded logic for decision making problems," *arXiv preprint arXiv:2505.14510*, 2025.

[9] C. De Silva, T. Halloluwa, and D. Vyas,"A multi-layered research framework for human-centered AI: Defining the path to explainability and trust," *arXiv preprint arXiv:2504.13926*, 2025.

[10] S. Hong, S. Fu, O. Serban, B. Bao, J. Kinross, F. Toni, G. Martin, and U. Vaghela,"Safeguarding patient trust in the age of AI: Tackling health misinformation with explainable AI," *arXiv preprint arXiv:2509.04052*, 2025.