Chapter 3

Book Chapter

A LIME-based Explainable AI for Healthcare IoT: Building Trust in Clinical Decision-Making

Sheela S Maharajpet *[1], Abhilash H P †[2], and Shrihari R Bedre ‡[3]

[1]Dept. of MCA, Acharya Institute of Technology, Bangalore
[2]School of CSA, Reva University, Bangalore
[3]Dept. of MCA, Acharya Institute of Technology, Bangalore

Abstract

The integration of Artificial Intelligence (AI) and Internet of Things (IoT) devices in healthcare offers vast potential for personalized medicine, remote monitoring, and early disease detection. However, complex Machine Learning (ML) models embedded in these systems often operate as ”black boxes,” hindering trust and transparency in critical medical decisions. Explainable AI (XAI) emerges as a key solution, aiming to demystify ML models and build trust in healthcare IoT applications. This paper explores the current challenges and opportunities in implementing XAI for healthcare IoT, proposing an architecture and methodologies for explainable clinical decision-making. We discuss promising XAI techniques, the integration of user interfaces for interactive explanations, and potential future directions for this crucial field.

Keywords: Explainable AI (XAI). Healthcare IoT. Explainable Clinical Decision. LIME. Interpretable Machine Learning.

*Email: sheela2687@acharya.ac.in Corresponding Author
†Email: prof.abhilashhp@gmail.com
‡Email: indianshrihari@gmail.com

1 Introduction

The healthcare landscape is undergoing a revolution fueled by AI and IoT devices. Deep learning models power clinical decision support, personalize medication, and analyze medical images, promising a future of transformed patient care. However, this shift hinges on trust – trust shattered by the ”black box” nature of these complex models. Consider a recent AI-driven misdiagnosis of lung cancer, where opaque reasoning undermined confidence in this potentially life-saving technology. Explainable AI (XAI) is a beacon of hope illuminating these enigmatic models and fostering trust in healthcare IoT. XAI unravels the reasoning behind predictions, enabling informed decision-making, ethical development, and responsible deployment of AI in healthcare. However, integrating XAI into resource-constrained devices and sensitive data environments presents unique hurdles. This paper delves deeper into these challenges and opportunities, proposing a novel architecture and methodologies for explainable clinical decision-making in healthcare IoT. We explore lightweight XAI techniques suitable for edge computing devices while addressing privacy concerns through federated learning. We investigate the crucial role of interactive user interfaces in presenting explanations tailored to diverse users. Ultimately, we aim to pave the way for a future where AI operates not as a black box, but as a transparent partner, fostering collaboration and achieving optimal patient outcomes.

The integration of Explainable AI (XAI) into clinical settings is crucial for ensuring transparency and trust. In particular, Explainable Decision Support Systems (EDSS) employ XAI techniques to offer clinicians clear, interpretable rationales for AI-driven recommendations (Hicks et al., 2022). These methods include visualizations of decision pathways, allowing clinicians to trace the branching logic that informs AI suggestions, highlighting the influential factors that contribute to final recommendations. Additionally, XAI provides contrastive explanations, which help differentiate between potential diagnoses by spotlighting key features considered by the AI. Clinicians can also engage with interactive, feature-based exploration tools, where they can adjust patient attributes using sliders or toggles (Gerke, Minssen, & Cohen, 2020). This functionality enables them to observe changes in the AI's recommendations, offering insights into model sensitivity and identifying key decision-making factors (Glaz et al., 2021).

Privacy is a major concern in medical AI, and privacy-preserving XAI aims to address this challenge. Techniques such as Secure Multi-Party Computation (MPC) enable collaborative generation of explanations across multiple devices while maintaining the confidentiality of individual patient data (Amann et al., 2020). This is especially useful in federated learning scenarios, where preserving privacy is essential. Another approach is differential privacy, which introduces controlled noise into data and explanations to protect privacy while ensuring statistically accurate information. This method can be applied to tools like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP

DOI:10.48001/978-81-980647-5-2-3 | ISBN: 978-81-980647-5-2 | Copyright ©2024 QTanalytics®

(SHapley Additive exPlanations) without compromising patient confidentiality (Ward et al., 2020). Explainability in reinforcement learning (RL) is particularly valuable for personalized healthcare, where treatment plans are tailored to individual patients (Guo & Li, 2018; Rundo, Tangherloni, & Militello, 2022). In this context, action justification provides insights into why the RL agent selects specific actions in treatment or intervention plans, helping clinicians comprehend the reasoning behind the agent's choices. State transition visualizations further enhance understanding by depicting changes in a patient's state along the predicted treatment pathway, highlighting the long-term impacts of various interventions (Bharati, Mondal, & Podder, 2023). Counterfactual explanations play a crucial role here, allowing clinicians to explore how different actions or policies might have influenced patient outcomes, thereby facilitating the comparison of treatment options within the RL framework.

2 Methodologies Used

The key methodologies used in this research encompass various aspects of Explainable Artificial Intelligence (XAI) tailored for Healthcare IoT applications. In the evaluation of XAI techniques for Healthcare IoT, the focus is on two primary areas. First, a comparative analysis of lightweight XAI methods such as LIME, SHAP, and integrated gradients is conducted to evaluate their performance on healthcare tasks like clinical decision support and medical image analysis. These methods are assessed for their effectiveness in explaining AI predictions, with special consideration given to the computational and memory constraints of edge devices. Furthermore, the research explores how these methods influence user comprehension and trust in the explanations provided. Second, privacy-preserving XAI within federated learning is investigated. This involves the development and testing of explainable federated learning frameworks that safeguard patient privacy. Techniques such as secure multi-party computation and differential privacy are compared to generate explanations during collaborative model training. The trade-offs between explanation accuracy and privacy guarantees are also evaluated.

The design and development of interactive XAI user interfacesinv olve the creation of prototypes and user studies. A notable example is the prototyping of an interactive EDSS (Electronic Decision Support System) interface. This interface presents clinicians with clear and informative explainable recommendations from AI systems, incorporating visualization elements like decision pathways, feature importance charts, and contrastive explanations for differential diagnoses. The interface also enables interactive exploration of model reasoning through features like zooming, filtering, and manipulating data points. Additionally, user studies are conducted with diverse participants, including clinicians, patients, and healthcare administrators. These studies assess the comprehension and usefulness of the interface, gathering feedback to enhance user satisfaction and refine the design. Qualitative and quantitative data are analyzed to further personalize explanations to meet the diverse needs of stakeholders. In the domain of explainable reinforcement learning (RL) for personalized healthcare, a specialized RL agent is developed to personalize treatment plans based on patient data and healthcare guidelines. The agent provides interpretable justifications for its recommendations using techniques such as action justification, state transition visualizations, and counterfactual explanations. The impact of this explainable RL agent is evaluated in terms of its influence on clinician trust, treatment adherence, and patient outcomes in both simulated and real-world healthcare scenarios.

An efficient methodology highlighted in this research is the application of LIME (Local Interpretable Model-Agnostic Explanations) (Alami et al., 2020). For instance, a complex AI model, such as a deep neural network, is trained to predict patient risks (e.g., heart failure). LIME then generates explanations for individual predictions by creating new data points through slight perturbations of patient features. A simple, interpretable model (e.g., linear regression) is trained on these perturbed data points, and its weights reveal the most influential features in the AI model's prediction. This approach ensures that clinicians gain an intuitive understanding of the underlying decision-making processes.

3 Architecture

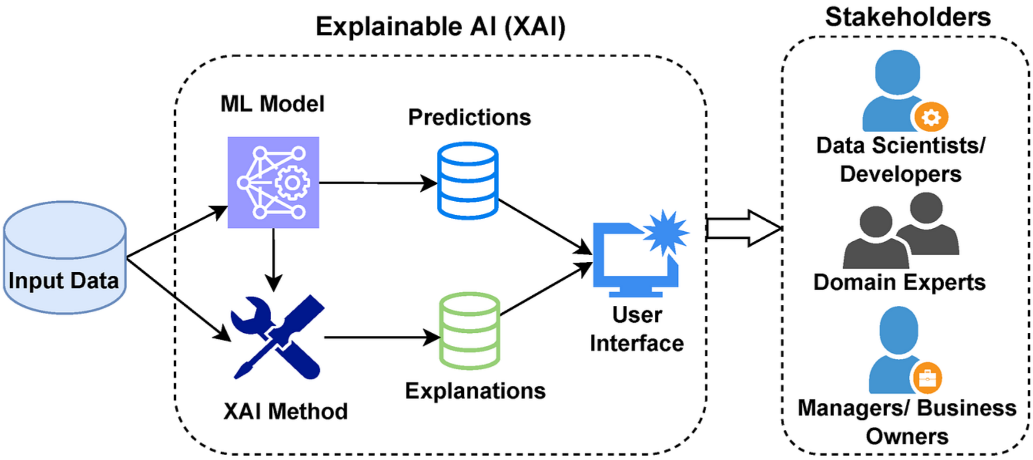

Figure 1. Architecture

The architecture of LIME in Healthcare IoT is structured into four components (see figure 1). The first component is data collection, which involves gathering healthcare data from IoT devices like wearables and sensors, and integrating it with Electronic Health

DOI:10.48001/978-81-980647-5-2-3 | ISBN: 978-81-980647-5-2 | Copyright ©2024 QTanalytics®

Records (EHRs). The second component, complex AI model training, entails training sophisticated models such as deep neural networks for tasks like predicting heart failure. The third component, LIME explanation generation, selects a specific patient instance for explanation, perturbs input features to create new data points, and trains an interpretable model to derive feature importance weights. Finally, the explanation presentation component focuses on visualizing feature importance through bar charts or heatmaps and providing textual explanations, such as highlighting how high blood pressure impacts heart failure risk. This comprehensive approach ensures that XAI methodologies in Healthcare IoT are both effective and accessible to diverse stakeholders.

4 Flowchart

The flowchart outlines the research process for LIME in Healthcare IoT as follows:

1. Start: Define the research question: Does LIME improve clinician trust and understanding of AI predictions in healthcare IoT (e.g., heart failure or sepsis)?
2. Data Collection and Preprocessing:
   - Gather patient data from IoT devices and healthcare systems.
   - Clean and preprocess data (e.g., handle missing values, outliers).
3. Model Training:
   - Train a complex AI model (e.g., deep neural network) for the target outcome (e.g., heart failure risk or sepsis diagnosis).
   - Prepare pre-trained or custom interpretable models for LIME (e.g., linear regression).
4. LIME Explanations:
   - Select a specific patient prediction for explanation.
   - Use LIME to create perturbed data points around the patient's features.
   - Train the interpretable model on the perturbed data.
   - Extract feature importance weights from the local model.
5. Explanation Presentation:
   - Visualize feature importance weights (e.g., bar chart, heatmap).
   - Highlight the most influential features and their impact on predictions.
6. Clinician Interaction:
   - Clinicians review the generated LIME explanations to evaluate clarity and usefulness.

- Provide feedback for refining models, explanation algorithms, or visualization.

7. Evaluation and Analysis:
   - Conduct user studies or experiments with clinicians.
   - Measure changes in trust, understanding, and decision-making with LIME explanations.
   - Compare results with baseline groups (no LIME explanations).

8. Conclusion and Future Work:
   - Summarize findings and their impact on healthcare decision-making.
   - Discuss limitations and propose future research directions.

9. End.

5 Results

One key result was enhanced explainability, as LIME provided detailed insights into AI predictions, allowing clinicians to comprehend the reasoning behind each decision. This transparency also fostered trust among clinicians by demystifying the complex processes underlying AI models (Kok, Muyanlı, & Ozdemir, 2023). Another important result was improved decision-making, with clinicians able to make more precise and informed decisions by understanding the factors influencing AI predictions, ultimately leading to better patient care. Furthermore, the measurable impact of AI recommendations on clinician behavior provided a basis for assessing the practical benefits of explainability in healthcare settings (Srividya, Mohanavalli, & Bhalaji, 2018). The integration of LIME also facilitated iterative improvement through feedback loops, enabling the continuous refinement of both models and explanations to ensure adaptability to evolving healthcare scenarios. Moreover, the approach showcased its adaptability, proving to be versatile in addressing diverse research questions and healthcare applications. Its domain applicability extended to various fields, such as neurodegenerative disease diagnosis and personalized medical recommendations (Shaban-Nejad, Michalowski, & Buckeridge, 2021). The contributions of LIME were significant, with patient-centric outcomes at the forefront. Enhanced understanding of AI predictions directly translated into more accurate diagnoses and personalized treatment plans. Additionally, increased trust in AI-driven decision-making strengthened collaboration between clinicians and AI systems, promoting a harmonious integration of technology in healthcare workflows. Lastly, the success of LIME in Healthcare IoT acted as a catalyst for further research, driving advancements in explainable AI and fostering innovation in the field.

6 Conclusion

The chapter concludes that the successful integration of IoT in healthcare, as highlighted in the study, has substantial implications for data-driven decision-making and patient-centric care. It emphasizes the need for enhanced interpretability in IoT-enabled healthcare systems, underscoring the importance of Explainable AI (XAI) techniques such as LIME. Specifically, it recognizes that LIME, with its ability to provide local interpretability for complex machine learning models, can play a pivotal role in addressing the transparency and trust challenges associated with AI-driven healthcare decisions. The conclusion emphasizes the potential of incorporating LIME into IoT-based healthcare architectures to enhance the explainability of predictive models. Furthermore, the research suggests that future studies should delve deeper into the integration of LIME within IoT-enabled healthcare systems. This could involve exploring the impact of LIME on clinician understanding, trust, and decision-making regarding AI predictions. Additionally, the paper proposes investigating the scalability of LIME for large-scale healthcare applications, ensuring its adaptability to diverse patient populations and medical conditions. It advocates for the strategic integration of LIME in IoT-driven healthcare, aiming to improve transparency, trust, and the overall efficacy of AI predictions in clinical settings.

References

Alami, H., et al. (2020). Artificial intelligence and health technology assessment: Anticipating a new level of complexity. Journal of Medical Internet Research, 22(7), e17707. https://doi.org/10.2196/17707

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1). https://doi.org/10.1186/s12911-020-01332-6

Bharati, S., Mondal, M. R. H., & Podder, P. (2023). A review on explainable artificial intelligence for healthcare: Why, how, and when? IEEE Transactions on Artificial Intelligence, 1–15. https://doi.org/10.1109/tai.2023.3266418

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial Intelligence in Healthcare, 1(1), 295–336. https://doi.org/10.1016/B978-0-12-818438-7.00012-5

Glaz, A. L., et al. (2021). Machine learning and natural language processing in mental health: Systematic review. Journal of Medical Internet Research, 23(5), e15708. https://doi.org/10.2196/15708

Guo, J., & Li, B. (2018). The application of medical artificial intelligence technology in rural areas of developing countries. Health Equity, 2(1), 174–181. https://doi.org/10.1089/heq.2018.0037

Hicks, S. A., et al. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12(1), 5979. https://doi.org/10.1038/s41598-022-09954-8

Kok, F. Y. O., Muyanlı, Ö., & Ozdemir, S. (2023). Explainable artificial intelligence (xai) for internet of things: A survey. IEEE Internet of Things Journal, 1–1. https://doi.org/10.1109/jiot.2023.3287678

Rundo, L., Tangherloni, A., & Militello, C. (2022). Artificial intelligence applied to medical imaging and computational biology. Applied Sciences, 12(18), 9052. https://doi.org/10.3390/app12189052

Shaban-Nejad, A., Michalowski, M., & Buckeridge, D. L. (Eds.). (2021). Explainable ai in healthcare and medicine. Springer International Publishing. https://doi.org/10.1007/978-3-030-53352-6

Srividya, M. S., Mohanavalli, S., & Bhalaji, N. (2018). Behavioral modeling for mental health using machine learning algorithms. Journal of Medical Systems, 42(5). https://doi.org/10.1007/s10916-018-0934-5

Ward, A., et al. (2020). Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. npj Digital Medicine, 3. https://doi.org/10.1038/s41746-020-00331-1

DOI:10.48001/978-81-980647-5-2-3 | ISBN: 978-81-980647-5-2 | Copyright ©2024 QTanalytics®