جامعة الأمير محمد بن فهد الأهلية
PRINCE MOHAMMAD BIN FAHD UNIVERSITY

A comprehensive framework for transparent and explainable AI sensors in healthcare

11th International Electronic Conference on Sensors and Applications

Part of the International Electronic Conference on Sensors and Applications series

26–28 Nov. 2024

Rabaï Bouderhem,
Assistant Professor
College of Law, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia
Research Associate, CREDIMI FRE 2003 CNRS – Université de Bourgogne, Dijon, France

1. Introduction

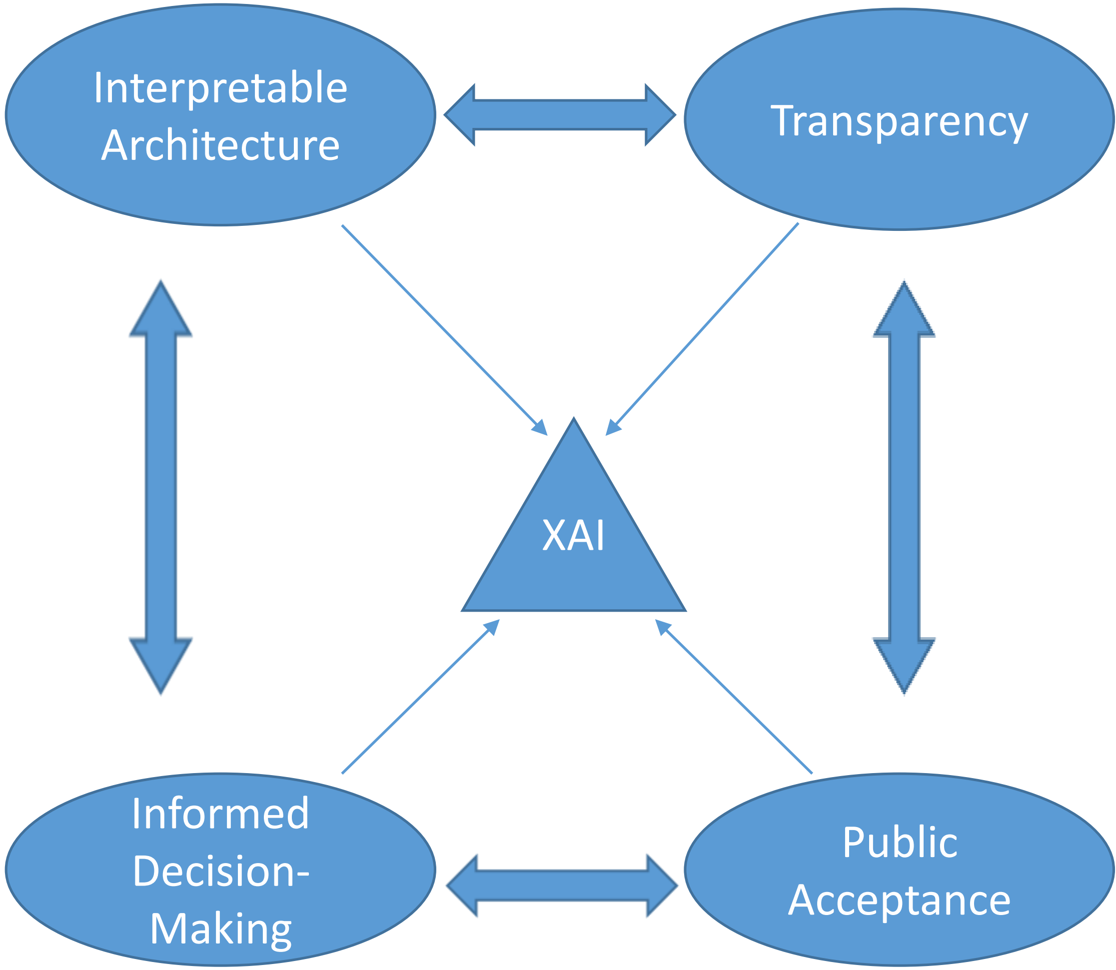
- The opaque nature of many current AI models, often referred to as "*black boxes*" [1], poses significant challenges in terms of interpretability, fairness, and reliability, which are critical factors in healthcare applications [2].
- The need for explainable and transparent AI (XAI) in healthcare has been widely acknowledged by researchers, practitioners, and policymakers.
- XAI aims to develop AI systems that are not only accurate and efficient [3] but also capable of providing human-understandable explanations for their decisions [4].
- By making AI systems more interpretable and transparent, XAI can foster trust [5], enable effective human-AI collaboration, and facilitate the responsible deployment of AI in healthcare [6].
- This research aims to address the challenges of developing explainable and transparent AI sensors for healthcare applications.
- Specifically, we propose a comprehensive framework that integrates interpretable machine learning models, human-AI interaction mechanisms, and ethical guidelines to ensure that AI sensor outputs are comprehensible, auditable, and aligned with clinical decision-making processes.

The proposed framework has ***three core components***:

- Firstly, an interpretable AI model architecture that leverages techniques such as attention mechanisms [7], symbolic reasoning [8], and rule-based systems [9] to provide human-understandable explanations.

- Secondly, an interactive interface that facilitates effective communication and collaboration between healthcare professionals and AI systems [10], enabling seamless integration of AI insights into clinical workflows.

- Thirdly, a robust ethical and regulatory framework that addresses issues of bias [11], privacy [12], and accountability [13] in the deployment of AI sensors in healthcare.



By developing explainable and transparent AI sensors tailored for healthcare applications, this research aims to contribute to the responsible development of AI technologies and pave the way for improved patient outcomes, informed decision-making, and increased public acceptance of AI in the healthcare domain [14].

2. Methodology

To develop a comprehensive framework for explainable and transparent AI sensors in healthcare, we employ a multi-pronged approach involving a systematic literature review and empirical analysis.

2.1. Comprehensive Literature Review

We conducted a comprehensive review of existing literature to identify the key requirements, challenges, and state-of-the-art techniques associated with developing transparent and explainable AI systems for healthcare applications.

We identified the critical factors for deploying AI systems in healthcare, such as interpretability [15], transparency, fairness, privacy, and accountability [16].

In addition, we examined the challenges and pitfalls of applying opaque "black-box" AI models in high-stakes healthcare situations [17].

We explored various interpretable machine learning models and techniques, including attention mechanisms, symbolic reasoning, and rule-based systems.

Then, we investigated human-AI interaction approaches for effective communication and collaboration between healthcare professionals and AI systems [18].

Finally, we analyzed ethical frameworks, guidelines, and regulatory considerations (HIPAA, AI Act, Data Act, GDPR…) for responsible AI deployment in healthcare [19].

2.2. Empirical Analysis

To validate and refine our proposed framework, an empirical analysis involving data collection, preprocessing, and experimental evaluation is necessary and should consist of the following steps:

1. Data Collection and Preprocessing

First, we need to gather relevant healthcare datasets (e.g., electronic health records, sensor data, and medical images) from publicly available sources or collaborating healthcare institutions. PubMed, Web of Science and Scopus databases could also serve as a starting point to collect relevant data. Second, we should preprocess the data to handle missing values, noise, and other data quality issues, while ensuring compliance with privacy and ethical guidelines.

2. Experimental Setup and Evaluation Metrics

The first step here is to implement and evaluate the components of our proposed framework, including interpretable AI models, interactive interfaces, ethical and regulatory considerations. The second step is to define appropriate evaluation metrics to assess the performance, interpretability, and transparency of our approach, such as predictive accuracy, model complexity, human-interpretability scores, and fairness measures so we can ensure data accuracy and relevance. The third step is to conduct controlled experiments and simulations to compare our framework with existing baseline methods and approaches.

3. Proposed Framework

- Building upon the insights gained from the literature review, we propose a comprehensive framework for developing explainable and transparent AI sensors in healthcare settings. The proposed framework consists of three core components:

1. Interpretable AI Model Architecture

2. Interactive Human-AI Interface

3. Ethical and Regulatory Framework

Objectives:
Development and deployment of explainable and transparent AI sensors in healthcare settings, fostering trust, accountability, and responsible AI adoption

Key aspects of the interpretable AI model architecture:
1. Attention mechanisms [20]
2. Symbolic reasoning [21, 22]
3. Rule-based systems [23]
4.Human-understandable explanations [24, 25, 26]

Key elements to be incorporated in the interactive human-AI interface:
1. Explanation visualization [27]
2. Interactive querying [28]
3. Collaborative workflow integration [29]
4. User feedback and model refinement [30]

Key ethical and regulatory challenges:
1. Bias mitigation, discrimination and fairness [31, 32, 33]
2. Privacy and data protection [34, 35, 36]
3. Accountability and auditing
4. Ethical guidelines and oversight
5. Transparency
6. Explainability
7. Performance [37]
8. Data quality and accuracy
9. Cost-effectiveness and affordability
10. Errors and misdiagnosis
11. Access to health and technology for all

4. Discussion and Future Directions

Enhancing Transparency, Explainability, and Trust
1. Interpretability of AI Sensor Outputs
2. Healthcare Professional-AI Collaboration
3. Addressing Ethical and Regulatory Concerns

- Our interpretable AI model architecture will have the ability to provide human-understandable explanations for AI sensor outputs, enhancing transparency and facilitating trust between healthcare professionals and AI systems [38].
- The interactive human-AI interface will facilitate effective communication and collaboration between healthcare professionals and AI systems, enabling a seamless integration of AI sensor insights into clinical workflows [39].
- Our ethical and regulatory framework will effectively mitigate biases in AI sensor outputs, reducing the risk of unfair treatment or discrimination against certain patient groups [40].
- Strong privacy-preserving measures and data protection techniques will ensure compliance with relevant regulations and protected sensitive patient data from potential privacy attacks or breaches.

Limitations and Future Research Directions
1. Scalability and Computational Complexity [41]
2. Generalizability across Healthcare Domains [42]
3. Continuous Model Refinement and Adaptation [43]
4. Integrating Multi-modal Data Sources [44]
5. Fostering Trust and Acceptance [45]

5. Conclusions

- The responsible development and deployment of AI technologies, particularly in high-stakes domains like healthcare, is of paramount importance.

- Our research contributes to this goal by providing a comprehensive framework that prioritizes transparency, explainability, and ethical considerations throughout the AI development lifecycle.

- By making AI systems more interpretable and facilitating human-AI collaboration, our approach empowers healthcare professionals to understand and trust the reasoning behind AI-driven recommendations and decisions.

- This trust is crucial for the successful adoption and integration of AI technologies in healthcare settings, ultimately contributing to improved patient outcomes and informed decision-making processes.

References:

1. Wadden JJ, Defining the undefinable: the black box problem in healthcare artificial intelligence, *Journal of Medical Ethics* 2022;**48:**764-768. https://doi.org/10.1136/medethics-2021-107529.

2. Valente, F., Paredes, S., Henriques, J. et al. Interpretability, personalization and reliability of a machine learning based clinical decision support system. Data Min Knowl Disc **36**, 1140–1173 (2022). https://doi.org/10.1007/s10618-022-00821-8.

3. Manresa-Yee, C., Roig-Maimó, M.F., Ramis, S., Mas-Sansó, R. (2022). Advances in XAI: Explanation Interfaces in Healthcare. In: Lim, CP., Chen, YW., Vaidya, A., Mahorkar, C., Jain, L.C. (eds) Handbook of Artificial Intelligence in Healthcare. Intelligent Systems Reference Library, vol 212. Springer, Cham. https://doi.org/10.1007/978-3-030-83620-7_15.

4. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of Explainable AI Techniques in Healthcare. Sensors **2023**, 23, 634. https://doi.org/10.3390/s23020634.

5. Gerlings, J., Jensen, M.S., Shollo, A. (2022). Explainable AI, But Explainable to Whom? An Exploratory Case Study of xAI in Healthcare. In: Lim, CP., Chen, YW., Vaidya, A., Mahorkar, C., Jain, L.C. (eds) Handbook of Artificial Intelligence in Healthcare. Intelligent Systems Reference Library, vol 212. Springer, Cham. https://doi.org/10.1007/978-3-030-83620-7_7.

6. Akhtar, M.A.K., Kumar, M., Nayyar, A. (2024). Socially Responsible Applications of Explainable AI. In: Towards Ethical and Socially Responsible Explainable AI. Studies in Systems, Decision and Control, vol 551. Springer, Cham. https://doi.org/10.1007/978-3-031-66489-2_9.

7. Rajabi, E.; Kafaie, S. Knowledge Graphs and Explainable AI in Healthcare. Information **2022**, 13, 459. https://doi.org/10.3390/info13100459.

8. Van Woensel, W., Scioscia, F., Loseto, G., Seneviratne, O., Patton, E., Abidi, S. (2024). Explanations of Symbolic Reasoning to Effect Patient Persuasion and Education. In: Juarez, J.M., et al. Explainable Artificial Intelligence and Process Mining Applications for Healthcare. XAI-Healthcare PM4H 2023 2023. Communications in Computer and Information Science, vol 2020. Springer, Cham. https://doi.org/10.1007/978-3-031-54303-6_7.

9. Kim, S.Y.; Kim, D.H.; Kim, M.J.; Ko, H.J.; Jeong, O.R. XAI-Based Clinical Decision Support Systems: A Systematic Review. Appl. Sci. **2024**, 14, 6638. https://doi.org/10.3390/app14156638.

10. Petersson, L., Larsson, I., Nygren, J.M. et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res **22**, 850 (2022). https://doi.org/10.1186/s12913-022-08215-8.

11. Angeliki Kerasidou, Ethics of artificial intelligence in global health: Explainability, algorithmic bias and trust, Journal of Oral Biology and Craniofacial Research, Volume 11, Issue 4, 2021, Pages 612-614, ISSN 2212-4268, https://doi.org/10.1016/j.jobcr.2021.09.004.

12. Nazish Khalid, Adnan Qayyum, Muhammad Bilal, Ala Al-Fuqaha, Junaid Qadir, Privacy-preserving artificial intelligence in healthcare: Techniques and applications, Computers in Biology and Medicine, Volume 158, 2023, 106848, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2023.106848.

13. A. Shaban-Nejad, M. Michalowski, J. S. Brownstein and D. L. Buckeridge, "Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2374-2375, July 2021, doi: 10.1109/JBHI.2021.3088832.

14. Kaifeng Liu, Da Tao, The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services, Computers in Human Behavior, Volume 127, 2022, 107026, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2021.107026.

15. MacDonald, S., Steven, K., Trzaskowski, M. (2022). Interpretable AI in Healthcare: Enhancing Fairness, Safety, and Trust. In: Raz, M., Nguyen, T.C., Loh, E. (eds) Artificial Intelligence in Medicine. Springer, Singapore. https://doi.org/10.1007/978-981-19-1223-8_11.

16. Ueda, D., Kakinuma, T., Fujita, S. et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol **42**, 3–15 (2024). https://doi.org/10.1007/s11604-023-01474-3.

17. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell **1**, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x.

18. Abedin, B., Meske, C., Junglas, I. et al. Designing and Managing Human-AI Interactions. Inf Syst Front **24**, 691–697 (2022). https://doi.org/10.1007/s10796-022-10313-1.

19. Haytham Siala, Yichuan Wang, SHIFTing artificial intelligence to be responsible in healthcare: A systematic review, Social Science & Medicine, Volume 296, 2022, 114782, ISSN 0277-9536, https://doi.org/10.1016/j.socscimed.2022.114782.

20. Park, S., Koh, Y., Jeon, H. et al. Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. Sci Rep **10**, 13413 (2020). https://doi.org/10.1038/s41598-020-70218-4.

21. Calegari, R.; Ciatto, G.; Denti, E.; Omicini, A. Logic-Based Technologies for Intelligent Systems: State of the Art and Perspectives. Information **2020**, 11, 167. https://doi.org/10.3390/info11030167.

22. Lu, Z., Afridi, I., Kang, H.J. et al. Surveying neuro-symbolic approaches for reliable artificial intelligence of things. J Reliable Intell Environ **10**, 257–279 (2024). https://doi.org/10.1007/s40860-024-00231-1.

23. Hassija, V., Chamola, V., Mahapatra, A. et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn Comput **16**, 45–74 (2024). https://doi.org/10.1007/s12559-023-10179-8.

24. Shahab S Band, Atefeh Yarahmadi, Chung-Chian Hsu, Meghdad Biyari, Mehdi Sookhak, Rasoul Ameri, Iman Dehzangi, Anthony Theodore Chronopoulos, Huey-Wen Liang, Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods, Informatics in Medicine Unlocked, Volume 40, 2023, 101286, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2023.101286.

25. Guleria, P., Srinivasu, P.N. & Hassaballah, M. Diabetes prediction using Shapley additive explanations and DSaaS over machine learning classifiers: a novel healthcare paradigm. Multimed Tools Appl **83**, 40677–40712 (2024). https://doi.org/10.1007/s11042-023-17212-w.

26. Juan M. Durán, Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare, Artificial Intelligence, Volume 297, 2021, 103498, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2021.103498.

27. Ooge, J., Stiglic, G., & Verbert, K. (2022). Explaining artificial intelligence with visual analytics in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1), e1427. https://doi.org/10.1002/widm.1427.

28. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput & Applic **32**, 18069–18083 (2020). https://doi.org/10.1007/s00521-019-04051-w.

29. Elham Nasarian, Roohallah Alizadehsani, U.Rajendra Acharya, Kwok-Leung Tsui, Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework, Information Fusion, Volume 108, 2024, 102412, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2024.102412.

30. Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, Zhe He, Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, Journal of the American Medical Informatics Association, Volume 27, Issue 7, July 2020, Pages 1173–1185, https://doi.org/10.1093/jamia/ocaa053.

31. Ferrara, E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci **2024**, 6, 3. https://doi.org/10.3390/sci6010003.

32. Ntoutsi E, Fafalios P, Gadiraju U, et al. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining Knowl Discov*. 2020; 10:e1356. https://doi.org/10.1002/widm.1356.

33. Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. ACM J. Responsib. Comput. 1, 2, Article 11 (June 2024), 52 pages. https://doi.org/10.1145/3631326.

34. Giuffrè, M., Shung, D.L. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digit. Med. **6**, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3.

35. Ali, A.; Pasha, M.F.; Ali, J.; Fang, O.H.; Masud, M.; Jurcut, A.D.; Alzain, M.A. Deep Learning Based Homomorphic Secure Search-Able Encryption for Keyword Search in Blockchain Healthcare System: A Novel Approach to Cryptography. Sensors **2022**, 22, 528. https://doi.org/10.3390/s22020528.

36. Rahman, A., Hossain, M.S., Muhammad, G. et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. Cluster Comput **26**, 2271–2311 (2023). https://doi.org/10.1007/s10586-022-03658-4.

37. Falco, G., Shneiderman, B., Badger, J. et al. Governing AI safety through independent audits. Nat Mach Intell **3**, 566–571 (2021). https://doi.org/10.1038/s42256-021-00370-7.

38. Ahmed, M., Zubair, S. (2022). Explainable Artificial Intelligence in Sustainable Smart Healthcare. In: Ahmed, M., Islam, S.R., Anwar, A., Moustafa, N., Pathan, AS.K. (eds) Explainable Artificial Intelligence for Cyber Security. Studies in Computational Intelligence, vol 1025. Springer, Cham. https://doi.org/10.1007/978-3-030-96630-0_12.

39. Chen, E., Prakash, S., Janapa Reddi, V. et al. A framework for integrating artificial intelligence for clinical care with continuous therapeutic monitoring. Nat. Biomed. Eng (2023). https://doi.org/10.1038/s41551-023-01115-0.

40. Chen, R.J., Wang, J.J., Williamson, D.F.K. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat. Biomed. Eng **7**, 719–742 (2023). https://doi.org/10.1038/s41551-023-01056-8.

41. Sarina Aminizadeh, Arash Heidari, Mahshid Dehghan, Shiva Toumaj, Mahsa Rezaei, Nima Jafari Navimipour, Fabio Stroppa, Mehmet Unal, Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service, Artificial Intelligence in Medicine, Volume 149, 2024, 102779, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2024.102779.

42. Nikolaos Koutsouleris, Tobias U Hauser, Vasilisa Skvortsova, Munmun De Choudhury, From promise to practice: towards the realisation of AI-informed mental health care, The Lancet Digital Health, Volume 4, Issue 11, 2022, Pages e829-e840, ISSN 2589-7500, https://doi.org/10.1016/S2589-7500(22)00153-4.

43. Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing. 2024. Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability. In Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI (CAIN '24). Association for Computing Machinery, New York, NY, USA, 100–111. https://doi.org/10.1145/3644815.3644959.

44. Huang, SC., Pareek, A., Seyyedi, S. et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digit. Med. **3**, 136 (2020). https://doi.org/10.1038/s41746-020-00341-z.

45. Kim, S.D. Application and Challenges of the Technology Acceptance Model in Elderly Healthcare: Insights from ChatGPT. Technologies **2024**, 12, 68. https://doi.org/10.3390/technologies12050068.

Thank you for your attention!