Can We Explain AI?: Explainable AI in the Health Domain as Told Through Three European Commission-funded Projects

Lem Ngongalah
Trilateral Research Ltd
London, UK
e-mail: Lem.ngongalah@trilateralresearch.com

Robin Renwick
Trilateral Research Ltd
Waterford, Ireland
e-mail: Robin.renwick@trilateralresearch.com

Abstract— **Artificial Intelligence (AI) has revolutionised healthcare, offering advanced diagnostics, personalised treatments, and enhanced patient outcomes. As AI increasingly integrates into healthcare systems, the need for Explainable AI (XAI) becomes paramount to ensure transparent and ethical decision-making. The lack of transparency and interpretability in AI systems poses significant challenges in healthcare, potentially undermining trust and hindering adoption. Understanding and addressing the complexities of XAI in healthcare is crucial for fostering trust among stakeholders, improving patient care, and adhering to ethical principles. Previous efforts have highlighted the importance of XAI but often lacked comprehensive approaches for implementation in diverse healthcare settings. This article explores the integration of XAI in healthcare, focusing on insights from three European Commission-funded projects under Horizon 2020/Horizon Europe. These projects prioritize transparency, accountability, and accessibility, showcasing the potential of XAI enhanced decision-making. In addition, the papers recognize the limitations of XAI, such as the absence of standardized approaches and the difficulty of balancing AI complexity with transparency, emphasizing the need for continuous refinement and adaptation to ensure successful XAI integration across varied healthcare settings.**

Keywords- ***Artificial intelligence; healthcare; explainable AI; trustworthiness; transparency***

I. INTRODUCTION

Artificial Intelligence (AI) has emerged as a transformative force in healthcare, offering unparalleled potential in diagnostics, treatment personalisation, and patient outcomes [1]. AI refers to the development of software that can use human-defined objectives to generate outputs such as content, predictions, or decisions influencing the environments they interact with [2]. AI systems encompass a broad spectrum of capabilities, with the capacity to perform varied tasks including problem-solving, learning, speech and pattern recognition, image classification and decision-making [2].

Explainable AI (XAI) is a critical component in AI design, ensuring transparent and understandable decision-making processes, in line with the AI Act [2] and Trustworthy AI guidelines [3], the foundational pillars of ethical AI development. From a computer science perspective, XAI involves developing algorithms that can provide interpretable insights into AI outcomes. In layman terms, XAI aims to provide a layer of understandability to algorithmic decision-making.

Recent controversies surrounding AI have highlighted concerns about the ethical implications of opaque decision-making processes, particularly in the healthcare sector where trust and understanding are crucial [4]. The integration of XAI is therefore not only desirable, but indispensable – imperative for the development of ethical, responsible, patient-centric, and trustworthy AI applications. However, one critical aspect of this integration is the quality, accuracy, and reliability of input data utilised by AI platforms. Healthcare data often faces challenges such as incompleteness, bias, and inconsistency, which can significantly impact performance and reliability. Addressing these data quality challenges is essential to ensure the effectiveness and trustworthiness of AI-driven decision-making processes in healthcare, thereby reinforcing the significance of XAI in healthcare settings.

This article explores the integration of XAI in healthcare, drawing insights from three European Commission-funded projects. Section I introduces the concept of Explainable AI and its significance in healthcare. Section II presents three case studies, highlighting their approaches to integrating XAI. Section III discusses the challenges and opportunities in implementing XAI in healthcare settings. Section IV concludes by outlining future directions to advance XAI in healthcare, with a focus on scalability, adaptability, and technical refinement.

II. CASE STUDIES

The iToBoS project (Grant agreement number 965221, April 2021 – March 2025) [6] aims to create an AI diagnostic platform for early skin melanoma detection, using a novel total body scanner and a computer-aided clinical decision support system that integrates patients' clinical information, genetic and imaging data, and family medical history. Despite advancements in AI, existing solutions often lack comprehensive transparency, leaving users with limited insights into decision-making processes. The project aimed to bridge this gap by integrating patient data and family medical history into a unified platform, ensuring comprehensible and transparent AI-driven diagnostics. To enhance the explainability of its AI components, the consortium prioritized transparency and interpretability through ongoing meetings within specific XAI work packages, interviews with project partners, end-users and patients, and an ongoing impact assessment process. These

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-136-7

assessments addressed concerns related to autonomy, transparency, and clinical effectiveness, as well as key considerations regarding patient understanding, potential conflicts between AI and clinician opinions, and the importance of clinician training. AI explainability extended to providing comprehensive insights into deep regression models. This involved adapting local XAI methods like Layer-wise Relevance Propagation for regression tasks, and using global XAI solutions, such as Concept Relevance Propagation to illustrate prediction strategies.

The COVINFORM project (Grant agreement number 101016247, November 2020 – October 2023) [7] explored the impacts of the COVID-19 pandemic across the EU member states and the UK, employing AI components to develop a risk assessment dashboard. Existing solutions often lack robustness in capturing multifaceted dimensions of vulnerability, hindering effective decision-making in pandemic response strategies. The project developed a comprehensive risk assessment dashboard, integrating statistical techniques and domain expertise to provide interpretable insights into various dimensions of vulnerability across regions and demographics, including physical, economic, social and information vulnerability. To prioritize explainability, the dashboard featured informative pop-ups/info-boxes providing interpretation guidance for dashboard outputs, and links to original data sources and metadata. End-user engagement was pivotal, employing a co-design approach involving workshops, usability testing to align technical and user needs, and cognitive walkthroughs where practitioners explored the dashboard, assessed semantic legibility, and performed tasks aligned with credible success story criteria. Recommendations from each testing phase informed subsequent usability testing rounds, refining the dashboard interface, and enhancing features relevant to the functioning and outcomes of the AI models. This iterative refinement highlights the project's commitment to user-friendly, interpretable AI-driven risk assessment tools for effective decision- making.

PREPARE-Rehab (Grant agreement number 10086219, June 2023 – May 2026) [8] aims to advance rehabilitation care for patients with chronic non-communicable diseases, by developing personalized, data-driven, computational prediction and stratification tools to enhance decision-making in selecting optimal therapy strategies. While existing solutions offer advancements in personalized medicine, they often lack transparency and user-friendliness, posing challenges in adoption and integration into clinical workflows. The project plans to address these limitations by prioritizing clear language, user-friendly interfaces, and the incorporation of graphical representations and visualization tools to enhance understanding of AI predictions. Comprehensive training for healthcare professionals is also prioritized, with emphasis on plain language and visual aids to bridge the gap between technical processes and user understanding enabling healthcare professionals to understand the advantages and limitations of AI tools. The project aims to create models with transparent decision-making processes, contributing to overall model interpretability and facilitating seamless integration of AI support in healthcare settings.

III. DISCUSSION

The three case studies presented in this extended abstract demonstrate a collective commitment to enhancing AI explainability while prioritizing transparency, accountability, and accessibility for non-technical users. By employing co-creation methodologies, these studies seek to enhance overall trustworthiness and understandability by integrating diverse perspectives throughout the development lifecycle. However, limitations exist in scaling such solutions across diverse healthcare environments, necessitating ongoing refinement and adaptation. Additionally, constraints exist in balancing the complexity of AI models with the imperative for transparency and comprehensibility, thus requiring ongoing discussion.

One significant challenge highlighted in these case studies is the lack of standardised approaches in XAI. The absence of universally accepted definitions for terms such as 'explainable' or 'interpretable' in the AI context has resulted in diverse approaches reflecting varying perspectives. This diversity complicates communication within the AI community and impedes the development of cohesive frameworks for evaluating and implementing XAI methodologies. To address this challenge, there is a need to identify and focus on specific aspects of XAI that can be standardised. By breaking down the field into identifiable parts, researchers and practitioners can work towards establishing internationally agreed standards. For instance, standardisation efforts could focus on defining key components of explainability, such as model interpretability, transparency in algorithmic decision-making, and methods for communicating AI outputs to diverse stakeholders. Furthermore, ongoing dialogue and knowledge exchange are essential for developing consensus-driven understandings of key AI concepts. Collaborative efforts, such as those within ISO [10] and CENELEC [11] play a crucial role in facilitating communication and laying the groundwork for international standards in XAI. However, uncertainties remain regarding the scalability and adaptability of co-creation methodologies in diverse cultural, industrial, and regulatory contexts. Future research and case studies, beyond the scope of EU projects, are needed to explore the broader applicability and challenges associated with scaling co-creation XAI methodologies.

IV. CONCLUSION AND FUTURE WORK

The future of XAI presents both challenges and opportunities. International standards will provide a platform for harmonised governance, conformity, and risk assessment. By identifying and standardizing key components of XAI, researchers and practitioners can facilitate smoother communication and foster sustainable innovation aligned with ethical and societal values. However, achieving this vision requires interdisciplinary collaboration, continuous dialogue, and a concerted effort to navigate evolving XAI techniques in an era of rapid technological advancement.

In our future work, we aim to further explore the scalability and adaptability of XAI methodologies, particularly across diverse cultural, industrial, and regulatory contexts. Furthermore, we plan to explore alternative approaches to addressing the varied explainability requirements for diverse stakeholders within the healthcare domain. In addition, our work will investigate the technical intricacies of implementing XAI models, including refining existing methodologies and developing novel techniques to enhance the transparency and interpretability of AI systems. Through these efforts, we aim to contribute to the ongoing evolution and refinement of XAI, ultimately enhancing trust, accountability, and accessibility in healthcare AI decision-making.

REFERENCES

[1] N. Hoppe, R. C. Härting, and A. Rahmel, "Potential Benefits of Artificial Intelligence in Healthcare," In Artificial Intelligence and Machine Learning for Healthcare, pp. 225-249, Springer, Cham, 2023.

[2] Proposal for a Regulation of The European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, Art 3(1). Accessible at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206 [retrieved: 8 February, 2024]

[3] High-Level Expert Group on AI: Ethics Guidelines for Trustworthy Artificial Intelligence. Accessible at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [retrieved: 5 February, 2024]

[4] J. P. Richardson, C. Smith, S. Curtis, S. Watson, X. Zhu, B. Barry, et al., "Patient apprehensions about the use of artificial intelligence in healthcare," NPJ digital medicine, 4(1), p.140, 2021.

[5] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," Information Fusion, 99, p.101805, 2023.

[6] Intelligent Total Body Scanner for Early Detection of Melanoma (ITOBOS). Accessible at: https://itobos.eu [retrieved: 5 February, 2024]

[7] COronavirus Vulnerabilities and INFOrmation dynamics Research and Modelling (COVINFORM). Accessible at: https://www.covinform.eu [retrieved: 8 February, 2024]

[8] PREPARE-Rehab. Accessible at: https://prepare-rehab.eu [retrieved: 8 February, 2024]

[9] W. J. Baumol, "The free-market innovation machine: Analyzing the growth miracle of capitalism," Princeton university press, 2002.

[10] ISO/IEC JTC 1/SC 42 - Artificial intelligence, https://www.iso.org/committee/6794475.html [retrieved: 10 February, 2024]

[11] CEN-CENELEC JTC 21 'Artificial Intelligence', https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence [retrieved: 10 February, 2024]