XAI in Healthcare*

Gizem Gezici[1,*], Carlo Metta[2], Andrea Beretta[2], Roberto Pellungrini[1],
Salvatore Rinzivillo[2], Dino Pedreschi[3] and Fosca Giannotti[1]

[1] *Scuola Normale Superiore, Pisa, Italy*
[2] *ISTI-CNR, Pisa, Italy*
[3] *University of Pisa, Italy*

Abstract

The evolution of Explainable Artificial Intelligence (XAI) within healthcare represents a crucial turn towards more transparent, understandable, and patient-centric AI applications. The main objective is not only to increase the accuracy of AI models but also, and more importantly, to establish user trust in decision support systems through improving their interpretability. This extended abstract outlines the ongoing efforts and advancements of our lab addressing the challenges brought up by complex AI systems in healthcare domain. Currently, there are four main projects: Prostate Imaging Cancer AI, Liver Transplantation & Diabetes, Breast Cancer, and Doctor XAI, and ABELE.

Keywords

Explainable AI, healthcare, interpretability, user trust

1. Introduction

AI-assisted clinical decision support systems (CDSS) [1] have brought many opportunities, mainly by leading to a better diagnosis performance, predicting patient outcomes, and personalising treatment plans. Nonetheless, there have been growing concerns due to the opaque nature of the widely-used black-box algorithms in CDSS. Our lab's work is dedicated to advancing the interpretability of AI models through local explanation techniques, with the ultimate goal of making AI decisions more transparent and comprehensible to healthcare providers and patients similar to the study of a CDSS in vaccine hesitancy through leveraging XAI approaches to obtain valuable insights about public health [2].

2. Methodology

Our methodology on XAI in healthcare field combines AI technologies with healthcare domain knowledge, and the key methodologies in our research are model-agnostic local explainers

HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 10—14, 2024, Malmö, Sweden

*Corresponding author.

gizem.gezici@sns.it (G. Gezici); carlo.metta@isti.cnr.it (C. Metta); andrea.beretta@isti.cnr.it (A. Beretta); roberto.pellungrini@sns.it (R. Pellungrini); rinzivillo@isti.cnr.it (S. Rinzivillo); pedre@di.unipi.it (D. Pedreschi); fosca.giannotti@sns.it (F. Giannotti)

0000-0001-9782-575 (G. Gezici); 0000-0002-9325-8232 (C. Metta); 0000-0001-8531-9325 (A. Beretta); 0000-0003-3268-9271 (R. Pellungrini); 0000-0003-4404-4147 (S. Rinzivillo); 0000-0003-4801-3225 (D. Pedreschi); 0000-0003-3099-3835 (F. Giannotti)

for generating understandable and relevant explanations of model predictions on healthcare datasets. Among these approaches, the first local explainer which works for different input data types provides decision rules of influential factors and counterfactual rules. The second approach we use specifically works on image data and returns a set of exemplar and counter-exemplar images, as well as a saliency map. Lastly, apart from the aforementioned explainers, the third local explainer is ontology-based that works on multi-labeled sequential data. In addition to the local explainers, we use global feature attributions to understand the overall model behaviour.

The key methodologies used in the following projects are related to the LORE method proposed by Guidotti et al. [3]. LORE is a powerful framework for generating local and interpretable explanations for machine learning models. LORE utilizes a genetic algorithm to create a synthetic neighborhood, which serves as the basis for training a local interpretable predictor. This predictor captures the underlying logic of the model's decision-making process, enabling the derivation of meaningful explanations. One of the key characteristics of LORE is its ability to provide transparent and understandable explanations for individual predictions. By focusing on local interpretability, LORE aims to explain the reasoning behind a specific prediction rather than the overall behavior of the model. This makes it particularly useful in situations where interpretability at the instance level is crucial, such as in healthcare and finance.

The explanations consist of two main components. First, a decision rule is derived from the logic of the local interpretable predictor. This decision rule sheds light on the factors that influenced the model's decision, providing insights into the important features and their corresponding weights. This information helps in understanding the key drivers behind the prediction. Additionally, LORE produces a set of counterfactual rules as part of the explanation. These counterfactual rules suggest modifications to the instance's features that would lead to a different outcome. By providing actionable suggestions for changing the input variables, LORE enables users to explore what-if scenarios and understand how small changes can influence the model's predictions. The availability of the LORE framework, along with the accompanying code[1], facilitates its adoption and implementation in various domains. In next sections different research project are described. They leverage over LORE methodology from different point of views.

3. Current Projects

In this section, we briefly present our ongoing projects on XAI in healthcare by referring to the XAI methodologies mentioned above.

Prostate Imaging Cancer AI In this project, the dataset consists of T2-weighted and Apparent Diffusion Coefficient (ADC) MRI scans that were gathered in cooperation with the doctors in Prostate Cancer Unit. To enhance knowledge of prostate cancer diagnosis, we mainly leverage the local explainer that works on images to produce insightful justifications for intricate imaging analyses. The project will explore the novel field of cross-domain explanations between T2-weighted and ADC images. Through this approach, we seek to facilitate communication

[1] https://github.com/riccotti/LORE

between various imaging modalities and promote a more comprehensive, integrated understanding of prostate cancer diagnosis, ultimately leading to improved patient outcomes and management.

Liver Transplantation and Diabetes This project aims to establish an Explainable CDSS to investigate if there are some pre-liver transplantation (pre-ltx) patient characteristics that might affect the glycemic status (condition of diabetes), i.e. if a non-diabetic patient becomes pre-diabetic or diabetic after the liver transplantation, and also the survival of a given patient. In this project, we work in collaboration with doctors from the Diabetology Department and we employ the liver transplantation dataset that they collected. This tabular dataset includes 1468 patients, 470 of whom had liver transplants with follow-up data for one and five years after the operation. The proposed pipeline is composed of two main parts: i. classification model for the prediction tasks of diabetes and survival, ii. exploiting global and local XAI methods to explain the overall model behaviour and individual patient predictions respectively through pinpointing the impactful pre-ltx features.

Breast Cancer In this project, the dataset consists of public health records in collaboration with administrative institutions gathered through voluntary efforts and arranged into linked tables that can be accessed using the SAS statistical tool developed by North Carolina State University. Due to the size and complexity of the dataset, and the incomplete documentation, extracting information is challenging. To address this, an entity-relationship (ER) diagram has recently been developed as a conceptual schema to choose suitable columns and our discussions are ongoing to identify the main research questions to which we can answer with this particular dataset.

Doctor XAI: Explainer for Sequential Patient Data *Doctor XAI: an ontology-based approach to black-box sequential data classification explanations* [4] describes an ontology-based technique that aims to explain black-box predicting multi-labeled, sequential, ontology-linked data. Formal representations of knowledge called ontologies are used in the methodology to encapsulate concepts and relationships unique to a given domain. In order to forecast the next visit, the study concentrates on explaining Doctor AI [4], a multilabel classifier that uses a patient's clinical history as input.

This project aims to establish an explainable CDSS which takes the clinical history of a patient (sequential data) and predict the next visit with a multi-label classifier. Then, leveraging ontologies specifically the ICD-9 ontology on the MIMIC-III dataset [5] which contains de-identified health-related sequential data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Our experiments on the proposed pipeline showed promising results in terms of capturing domain-specific knowledge, extracting relevant features, and providing interpretable explanations. Currently, we further aim to refine and expand the capabilities of the proposed pipeline by utilizing Large Language Models (LLMs).

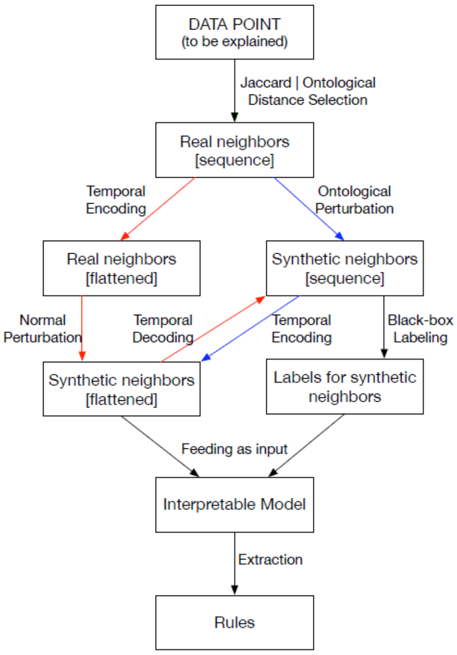

Figure 1: Doctor XAI explanation pipeline

ABELE ABELE (Adversarial Black-box Explainer generating Latent Exemplars) [6] is a local model-agnostic explainer that receives a picture as input, a black-box classifier, and sets of exemplar and counter-exemplar images along with a saliency map. Exemplars and counter-exemplars are artificially created images that are categorized with an outcome that differs from the input image and the same outcome as the input image, respectively. To comprehend the rationale for the choice, they can be visually examined. The input image's regions that support one class and those that force it into a different one are indicated by the saliency map. An Adversarial Autoencoder (AAE) is used by ABELE to create a neighborhood in the latent feature space. The encoder uses latent features to return the latent representation after receiving the image to be explained as input from the AAE. The neighborhood generation was achieved via a genetic technique that maximizes a fitness function. Utilizing a latent form of LORE, ABELE benefits in this way.

Following generation, ABELE queries the discriminator and converts the resultant image to verify the legitimacy of every instance within the neighborhood. Following that, it uses the picture to ask the black-box classifier for the class. By using the black-box classifier to label the neighborhood, ABELE constructs a decision tree classifier based on the local neighborhood. With the help of the surrogate tree, the black-box classifier's local behavior should be mimicked. The process facilitates the creation of exemplars and counter-exemplars by extracting the decision rule and counter-factual rules. The quality of the encoder and decoder functions used determines how effective ABELE is overall. The explanations will be more practical and meaningful the higher the AAE.

4. Conclusion

The vision of our lab on XAI in healthcare is on creating a powerful, accessible, and trustworthy AI-assisted CDSSs for healthcare professionals and patients. We believe that properly designing the integration of XAI methodologies based on the feedback from our healthcare practitioner collaborators is valuable. In this way, XAI can help us to foster trust, enhance decision-making, and improve treatments for patients. Going forward, the emphasis will continue to be on establishing CDSSs with AI in a manner that values human welfare above all else and acknowledges the intricacies of healthcare.

Acknowledgements

This work has been supported by the European Union under ERC-2018-ADG GA 834756 (XAI), by HumanE-AI-Net GA 952026, and by the Partnership Extended PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI”.

References

[1] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, C. Mooney, Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review, Applied Sciences 11 (2021) 5088.

[2] C. Punzi, A. Maslennikova, G. Gezici, R. Pellungrini, F. Giannotti, Explaining socio-demographic and behavioral patterns of vaccination against the swine flu (h1n1) pandemic, in: World Conference on Explainable Artificial Intelligence, Springer, 2023, pp. 621–635.

[3] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Factual and counterfactual explanations for black box decision making 34 (2019) 14–23.

[4] C. Panigutti, A. Perotti, D. Pedreschi, Doctor XAI: an ontology-based approach to black-box sequential data classification explanations, in: Conference on Fairness, Accountability, and Transparency, 2020, pp. 629–639.

[5] A. E. W. Johnson, MIMIC-III, a freely accessible critical care database, Scientific data 3 (2016).

[6] R. Guidotti, A. Monreale, S. Matwin, D. Pedreschi, Black box explanation by learning image exemplars in the latent feature space, in: U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, C. Robardet (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2020, pp. 189–205.