

Designing an evaluation framework for eXplainable AI in the Healthcare domain

Ivania Donoso-Guzmán^{1,2}

¹*KU Leuven, Department of Computer Science*

²*Pontificia Universidad Católica de Chile*

Abstract

The rapid adoption of Artificial Intelligence (AI) has brought automation and problem-solving capabilities in various fields, including healthcare. However, a significant challenge lies in the lack of explanation for AI predictions, particularly in healthcare, where transparency is crucial. This issue has led to eXplainable AI (XAI) development, focusing on constructing explanations for AI systems. However, the evaluation of these explanations lacks a standardized user-centric approach. This research proposes an evaluation framework for XAI methods to address this gap. The project involves four stages: conducting a systematic review of current evaluation methods, assessing the appropriateness of automatic evaluation of explanations, and conducting user studies to gauge the framework's effectiveness in capturing the user experience complexity. The desired outcome is a user-centric evaluation framework and guidelines, enhancing the scalability of XAI research and fostering confidence in adopting AI systems in the healthcare domain.

Keywords

Explainable AI, User-Centric evaluation, Human-Centered AI

1. Motivation

Research on artificial intelligence systems that seek to support medical teams in decision-making has shown outstanding progress in recent years. These systems aim to automate the greatest number of tasks and/or provide summarized and selected information to those who make decisions in a hospital environment [1]. This allows healthcare professionals to spend more time with patients on clinical tasks. However, these systems can make mistakes, have significant degrees of uncertainty [2], or even have important biases [3]. Since clinicians are responsible for making decisions that impact the well-being or life of people, this current state of technology makes it difficult to adopt these systems confidently [4, 5, 6]. In this context, it has been proposed that providing explanations about AI models or single predictions could potentially increase clinicians' trust [7] and ultimately boost adoption in healthcare settings [5].

EXplainable Artificial Intelligence (XAI) is the sub-area of Artificial Intelligence that aims to develop systems humans can understand, with the ultimate goal of increasing trust in AI systems. Research in XAI has mainly focused on developing algorithms that explain predictions.


Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

✉ indonoso@uc.cl (I. Donoso-Guzmán)

🆔 0000-0002-2427-9128 (I. Donoso-Guzmán)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Different taxonomies have been defined to classify them, and methods work with different data types (text, images, tabular data, time series). In the healthcare domain, some works have shown the potential of using XAI to uncover models' inner workings and explain single predictions. Zech et al. [8] used saliency maps to show the learned patterns of a chest X-ray CNN-based classifier. The model's goal was to predict the risk of pneumonia. The authors found that the CNN used non-disease-related image features to predict the risk. Kim et al. [9] presented an XAI system that connected visual features with appropriate semantic concepts to explain its predictions in diabetic retinopathy. Gutiérrez et al. [10] presented a user study on a call recommendation system for nursing homes. In this study, the healthcare professionals had to decide which call to attend based on the recommendation and explanations from the mobile application.

Although this area shows promising results, the value of explainable AI methods still has to be proven to work in practice [11] and this makes it difficult to deploy these systems in their respective domains [11, 12, 13, 6]. Methodologies from the AI/ML community, such as evaluation with a Ground Truth Dataset, cannot be used for these methods: the success of an explanation depends on the user, its context, the AI model and the explanation object. Since no clear consensus exists on properties and concepts related to explanations [14, 15], it has been challenging to define a formal evaluation procedure for XAI methods. In addition, while many researchers stress the importance of context, we are not aware of XAI evaluation methods that treat explanations' effects on humans as a **complex user experience** that assesses the effect of different explanation characteristics in the user experience. To better evaluate XAI methods, researchers have tried to disentangle explanation's characteristics into simple, measurable properties such as completeness [14, 16], novelty [17, 18], and interactivity [14, 16]. However, there is little evidence on how these properties relate to explanations being appropriate in real scenarios [17]. It has been proposed that some properties have relations with the user experience; for instance, simpler explanations generate more understanding [11, 17], but many of these relations have not been established with user studies.

In this work, we aim to discover what characteristics of explanations of AI predictions affect the user experience in terms of satisfaction, understanding, trust, reliability, adoption propensity and task performance. For this, we will design an evaluation framework for XAI-generated explanations in the context of healthcare applications with a user-centric approach. The healthcare domain has several characteristics that make it an excellent place to design a framework that could be later extended and generalized to other domains: AI technology has shown promising results, there is interest in the community in developing better and more robust tools, and there are different kinds of contexts (time to make a decision, user knowledge, interaction object) in which the XAI systems can be applied. Our main goal is to understand how explanations affect the user experience and to do so, we will propose an evaluation framework of XAI-based systems in the healthcare domain. This development could help to increase the deployment of AI applications in the domain, with the ultimate goal of improving healthcare practice.

2. Related work

2.1. Evaluation methods in XAI

Even though AI/ML models have standard evaluation metrics, there is still no consensus on the strategy to evaluate XAI methods. According to the study by Nauta et al. [14], 58% studies that present an XAI method performed a quantitative evaluation, 33% provided only anecdotal evidence, 17% performed a proxy task user study and only 5% performed a user study with domain experts. Within the studies that performed quantitative studies, they found no consistency in how the methods were evaluated. Seventy-three papers evaluated only one property of explanations, and six of the 12 properties were measured by very few studies.

According to Doshi-Velez and Kim [19], the metrics could be performed in three levels: application-grounded, with real tasks and users; human-grounded, with real users and proxy tasks; and functionality-grounded, with proxy tasks and no users. Most of the XAI evaluations use an application or human-grounded approach. These user studies have been criticized for their lack of rigor and for the use of proxy tasks [20].

To conduct functionally-grounded evaluations, i.e. proxy tasks and no users, some studies have focused on grouping concepts and defining properties [11, 16, 18] and their corresponding metrics [14]. These works aggregate existing literature that defines properties or presents metrics to assess them. The properties that have been proposed in these works try to measure the quality of the explanations without context so that they can be used in functionality-grounded evaluation. Recently [21] presented a framework to benchmark different XAI methods using automatic metrics. Still, it is limited to certain methods and only works with specific datasets created specifically for the benchmark.

2.2. AI and XAI in healthcare

The increasing availability of healthcare-related data has led to the creation of several applications of AI in the domain. From pattern detection using wearable data to using Electronic Health Records to improve patient care, Artificial Intelligence has started to play a role in the clinical research. However, few AI-based applications have been deployed in real clinical settings [13]. Some studies [12, 6] have suggested that the lack of trust of clinicians in the AI systems has a big impact on system deployments. There have been some [7, 1] qualitative studies to understand the interactions between medical staff and AI/ML systems, but they have focused on system aspects, not on the medical practice.

There is no consensus on the role of explanations in the clinical domain. Ghassemi et al. [22] states that explanations are not the right direction to increase the deployment of such systems. Instead, they propose using validation methods already in use in the medical community. On the contrary, Amann et al. [23] proposes to include explanations depending on the application context as long as they are proven to work.

3. Research Problem

3.1. Problem statement

The current situation of healthcare, AI, and XAI can be summarized as follows. Several AI models have proven to work for specific medical tasks. However, these models can be biased and always have uncertainty. In the healthcare domain, clinicians are responsible for decisions that affect people's lives; therefore, they need to ensure they understand what the AI models are predicting. One way to achieve this is to generate explanations of what the model is doing so that users can understand the logical procedure. Methods are being developed to generate explanations, but there are many definitions of explanations; therefore, it is complicated to describe what people could expect from them. This is because it is not well-defined what would be a good explanation in different contexts and for different users. Accordingly, it is challenging to establish a consistent evaluation that could increase adoption in the healthcare domain.

To tackle this problem, we proposed to first establish an evaluation procedure for XAI systems and, second, use this framework to understand how explanations affect the user experience in the healthcare domain. By deeply understanding these connections, guidelines for designing appropriate XAI systems can be created and used to develop better systems that increase adoption in the domain.

3.2. Objectives

This project has two goals: **(O1)** to develop a standard evaluation procedure and **(O2)** to generate guidelines for designing XAI systems. The first objective is decomposed into the following sub-goals

1. Define properties of explanation and their associated measurements.
2. Design, an evaluation framework for XAI-generated explanations that can be applied in the healthcare domain
3. Find relations between properties to explain the user experience
4. Provide researcher guidelines to help decide the properties and measurements more suited to the user study characteristics
5. Provide guidelines on user studies reporting to increase comparability between studies

Once the evaluation procedure has been proposed, we will be able to use it to answer the research questions by conducting user studies using the framework. With the outcome of these user studies, we will be able to accomplish the second objective of generating design guidelines.

3.3. Research questions

To be able to generate guidelines for XAI systems design we want to answer the following questions:

- (RQ1) What are the characteristics of explanations that relate to the user's experience?
- (RQ2) How do explanation characteristics affect the user experience in terms of satisfaction, trust, understanding, adoption, task performance and reliability? Are there mediation effects with other variables or characteristics that affect the user experience?
- (RQ3) What measurements or combination of those are more effective for measuring properties of explanations in the healthcare domain?
- (RQ4) How could these metrics be operationalized to measure explanations from models of different natures and for any dataset?

4. Research plan

To answer the research questions, I am working in four stages that depend on each other. The project is divided in two parts: first proposing

4.1. Stage 1: Taxonomy of explanations evaluation (RQ1, O1)

This stage will work towards achieving answering RQ1. During this stage, works in the area of evaluation of explanations will be reviewed to understand the following:

1. What are the properties of explanations according to current literature?
2. What are the metrics that are used to measure said properties?
3. What is the relationship between the properties and the user experience?
4. What are the measurement models that are used to evaluate the user experience?

4.2. Stage 2: Automation of evaluation (RQ4)

This stage aims to create procedures to measure certain metrics related to explanation aspects to answer RQ4. To accomplish this, we will create a Python package that implements the metrics that can be automated. This package could be used with any combination of XAI method, AI model, and dataset.

The metrics that will be implemented are currently being evaluated. Metrics that measure properties that have more connections with others and have been described as more relevant by previous authors will be prioritised. Additionally, we will conduct user studies to evaluate the alignment of these automatic metrics with the user experience. We will conduct these evaluations in domains where users can be found more easily than healthcare, such as recommendation systems, to conduct a quantitative study. This will consist of evaluating the user experience with questionnaires and applying the metrics to the examples shown to the user. Both measurements will be compared to understand whether the metric correctly measures the user experience property.

4.3. Stage 3: Evaluate alignment between theoretical properties and user experience (RQ2, O2)

During this phase, we will conduct qualitative user studies in a clinical setting to evaluate whether the properties of explanations we can measure align with the user experience. In this stage, we also want to understand the factors that affect users in a healthcare context. We will work in systems that area being developed by the Augment and HAIVis research groups. These projects work with image classification and interactive dashboards that help clinicians make decisions. The user study will be a semi-structured interview and analysed using thematic analysis. It considers a minimum of 5 participants, and we will stop recruiting when reaching saturation of themes.

The outcome of this stage will be an analysis of how the theoretical properties we found in the previous stage work in practice in a healthcare-related setting. This will help us to answer RQ2 and will contribute to O2 as well.

4.4. Stage 4: propose XAI evaluation guidelines (RQ2, RQ3, O2)

The study in stage 3 will allow us to understand more deeply what properties give value to the user experience. In this stage, we will test the complete evaluation framework quantitatively. We will use XAI applications in the healthcare domain, and we will apply the evaluation framework. The applications will be part of the current efforts of both research groups. The questionnaires for the user experience will be validated using confirmatory factor analysis to select the questions that more appropriately represent each property.

The user studies will shed light on the relevant properties of explanation for the medical domain, and we will be able to understand the relation between the properties of explanations in this particular area. This new understanding will help us to answer RQ1, which refers to the relation between the explanation properties in this particular domain. At the end of this stage, we will be able to achieve O2; We expect to provide guidelines for the development of XAI applications in the healthcare domain sustained in the evidence provided by the user studies that were conducted.

5. Contributions

The contributions of this work are the following: we will provide the research community with an evaluation procedure applicable to XAI systems based on evidence of multiple works in different domains, and by conducting user studies using this framework, we will be able to generate guidelines for XAI design based on empirical evidence that could be used to design more appropriate XAI systems.

Acknowledgments

This research is supported by ANID BECAS/DOCTORADO NACIONAL 21202228, Basal Funds for Center of Excellence FB210017 (CENIA), the Research Foundation Flanders (FWO, grant G0A3319N) and KU Leuven (grant C14/21/072).

References

- [1] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, M. Terry, Human-centered tools for coping with imperfect algorithms during medical decision-making, in: Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, 2019. doi:10.1145/3290605.3300234.
- [2] M. A. Ahmad, A. Teredesai, C. Eckert, Interpretable Machine Learning in Healthcare, in: 2018 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2018, pp. 447–447. URL: <https://ieeexplore.ieee.org/document/8419428/>. doi:10.1109/ICHI.2018.00095.
- [3] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, M. Ghassemi, Hurtful words, in: Proceedings of the ACM Conference on Health, Inference, and Learning, ACM, New York, NY, USA, 2020, pp. 110–120. URL: <https://dl.acm.org/doi/10.1145/3368555.3384448>. doi:10.1145/3368555.3384448.
- [4] F. Wang, R. Kaushal, D. Khullar, Should health care demand interpretable artificial intelligence or accept "black Box" Medicine?, 2020. doi:10.7326/M19-2548.
- [5] C. M. Cutillo, K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, K. D. Mandl, T. Beck, E. Collier, C. Colvis, K. Gersing, V. Gordon, R. Jensen, B. Shabestari, N. Southall, Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency, 2020. doi:10.1038/s41746-020-0254-2.
- [6] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, M. Ghassemi, Do as AI say: susceptibility in deployment of clinical decision-aids, *npj Digital Medicine* 4 (2021). URL: <http://dx.doi.org/10.1038/s41746-021-00385-9>. doi:10.1038/s41746-021-00385-9.
- [7] S. Tonekaboni, S. Joshi, M. D. Mccradden, A. Goldenberg, A. G. Ai, What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use, in: Proceedings of Machine Learning Research, PMLR, 2019, pp. 359–380. URL: <http://proceedings.mlr.press/v106/tonekaboni19a.html>.
- [8] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, E. K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, *PLOS Medicine* 15 (2018) e1002683. URL: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683>. doi:10.1371/JOURNAL.PMED.1002683.
- [9] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) (2018).
- [10] F. Gutiérrez, N. N. Htun, V. Vanden Abeele, R. De Croon, K. Verbert, Explaining Call Recommendations in Nursing Homes: a User-Centered Design Approach for Interacting with Knowledge-Based Health Decision Support Systems, *International Conference on Intelligent User Interfaces, Proceedings IUI (2022)* 162–172. URL: <https://dl.acm.org/doi/10.1145/3490099.3511158>. doi:10.1145/3490099.3511158.
- [11] A. F. Markus, J. A. Kors, P. R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design

- choices, and evaluation strategies, *Journal of Biomedical Informatics* 113 (2021) 103655. doi:10.1016/J.JBI.2020.103655.
- [12] I. A. Scott, S. M. Carter, E. Coiera, Exploring stakeholder attitudes towards AI in clinical practice, *BMJ Health & Care Informatics* 28 (2021) e100450. URL: <https://informatics.bmj.com/lookup/doi/10.1136/bmjhci-2021-100450>. doi:10.1136/bmjhci-2021-100450.
- [13] J. Shaw, F. Rudzicz, T. Jamieson, A. Goldfarb, Artificial Intelligence and the Implementation Challenge, *Journal of medical Internet research* 21 (2019). URL: <https://pubmed.ncbi.nlm.nih.gov/31293245/>. doi:10.2196/13659.
- [14] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI (2022). URL: <http://arxiv.org/abs/2201.08164>.
- [15] M. A. Clinciu, A. Eshghi, H. Hastie, A study of automatic metrics for the evaluation of natural language explanations, *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference* (2021) 2376–2387. doi:10.18653/v1/2021.eacl-main.202.
- [16] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion* 76 (2021) 89–106. doi:10.1016/J.INFFUS.2021.05.009.
- [17] Q. V. Liao, Y. Zhang, R. Luss, F. Doshi-Velez, A. Dhurandhar, Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI, *Proceedings of the AAI Conference on Human Computation and Crowdsourcing 10* (2022) 147–159. URL: <https://ojs.aaai.org/index.php/HCOMP/article/view/21995>. doi:10.1609/hcomp.v10i1.21995.
- [18] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, *Electronics (Switzerland)* 8 (2019) 1–34. doi:10.3390/electronics8080832.
- [19] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, *Arxiv* (2017) 1–13. URL: <http://arxiv.org/abs/1702.08608>.
- [20] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, *International Conference on Intelligent User Interfaces, Proceedings IUI* (2020) 454–464. doi:10.1145/3377325.3377498.
- [21] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, OpenXAI: Towards a Transparent Evaluation of Model Explanations, in: *NeurIPS, 2022*. URL: <https://arxiv.org/abs/2206.11104v2>. doi:10.48550/arxiv.2206.11104.
- [22] M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, *The Lancet Digital Health* 3 (2021) e745–e750. URL: <http://www.thelancet.com/article/S2589750021002089/fulltext>. doi:10.1016/S2589-7500(21)00208-9.
- [23] J. Amann, D. Vetter, S. N. Blomberg, H. C. Christensen, M. Coffee, S. Gerke, T. K. Gilbert, T. Hagendorff, S. Holm, M. Livne, A. Spezzatti, I. Strümke, R. V. Zicari, V. I. Madai, o. b. o. t. Z.-I. initiative, To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems, *PLOS Digital Health* 1 (2022) e0000016. doi:10.1371/JOURNAL.PDIG.0000016.