EXPLAINABLE AI IN MEDICINE

Holzinger et al. (2019)

Report

Explainable AI – IT790A

Lorenz Koch; Kim Wurster

A24lorko; A24kimwu

2025-10-12

Contents

Abstract

This report explores the role of Explainable Artificial Intelligence (XAI) in medicine, based on Holzinger et al. (2019), Causability and Explainability of Artificial Intelligence in Medicine. The paper emphasizes that clinical AI systems must provide explanations that are not only technically accurate but also meaningful to humans. It distinguishes between explainability, a property of the system, and causability, a measure of how well humans can derive causal understanding from explanations. Using the example of histopathology, it illustrates the gap between human causal reasoning and AI's statistical correlations. It concludes that medicine requires hybrid systems, combining the predictive strength of deep models with human-understandable reasoning to ensure trust, safety, and accountability in clinical decision-making.

1. Introduction

Explainable AI (XAI) is increasingly used in medical fields such as imaging, diagnostics, and decision support, where transparency and trust are essential. The Holzinger et al. (2019) paper focuses on histopathology imagery, where they describe how a pathologist annotates regions of interest and provides causal reasoning, which the AI then has to mimic or complement which aim is to ensure that the AI provides human understandable causal reasoning. This application shows that XAI is not just a theoretical concept but a practical necessity for bringing AI into clinical practice.

The paper we mainly focus on is Holzinger, Langs, Zatloukal & Müller (2019), "Causability and explainability of artificial intelligence in medicine" which was picked because it provides a good conceptual framework for aligning technical AI explanations with human reasoning in medicine. The central point is distinguishing explainability (what the AI system can show) from causability (how a human can derive causal insights).

2. Why does medicine need XAI?

XAI in medicine is important because medicine is a high stakes domain, where the decisions have direct consequences on the patient's health, safety and trust. Errors might have fatal results and misinterpretations of and by AI can lead to misdiagnosis, wrong treatment or even harm. Because of that, clinicians can't just blindly trust on a "black-box", they need to understand the output of the AI model to ensure that the results are acceptable. Holzinger et al. emphasizes that explainability is essential for clinical trust, safety and accountability. Clinicians must understand uncertainty and reasoning paths if the AI judgement is conflicting and because medicine often deals with uncertain and noisy data. Having only a probabilistic output is not enough as it might rely on unclear data inputs.

Decisions in the medical field are based on causal, mechanistic and domain knowledge, so AI models that only output correlations are of limited use. Holzinger et al. (2019) argue that explainable medicine demands not only that AI indicate what features contributed, but that it enables causal inference and counterfactuals (what-if scenarios). The explanations should support human causal reasoning and not just show the statistical associations.

Medicine needs AI whose internal reasoning can be investigated, understood and aligned with human causal models.

3. Human vs. Machine Explanations

Human clinicians explain their decisions through structured reasoning where they consider multiple interacting features. Their explanations usually follow causal patterns: “Because I see fibrosis in this region, and infiltration here, that suggests disease X, which would lead to symptom Y.” In the histopathology use case, the paper has a human annotator giving both post hoc (marking relevant regions) and ante hoc reasoning (the chain of logic that leads to diagnosis), showing how human reasoning works and that human explanations are more than feature lists, as they embed semantics, counterfactual thinking and domain knowledge which they have accumulated over years.

AI models explain via technical tools like attribution, saliency, heatmaps and other methods, which highlight what input parts contributed to a decision. These explanations however are often based on correlations found in the data and not causal. They don’t express the chain of reasoning or use medical concepts to make a diagnosis.

For instance, a saliency map may highlight a region without explaining why it is medically relevant as it is not providing that level of abstraction. Holzinger et al. emphasize that machine-generated explanations must be translated into human-understandable causal reasoning, which is one of the main gaps to cause practical challenges when using AI in medicine.

4. Glass-Box vs. Blackbox

Glass-box models or interpretable models are designed to be transparent by architecture such as decision trees, generalized additive models (GAMs) or others. Those models are structured in a way that is human readable, which enables people to trace decision paths or inspect weight contributions the models output and they are easier to audit, verify and understand, therefore providing less risk of hidden biases, as clinicians can inspect and challenge the model’s logic explicitly. Holzinger et al. (2019) states that there are models which have reached a competitive performance by focusing on the feature engineering and with the current state of XAI those models should preferably be used.

The negative side of models is that they often struggle with complex, high-dimensional raw data and often cannot compete with the performance of deep learning models, especially in vision related tasks such as analyzing scans.

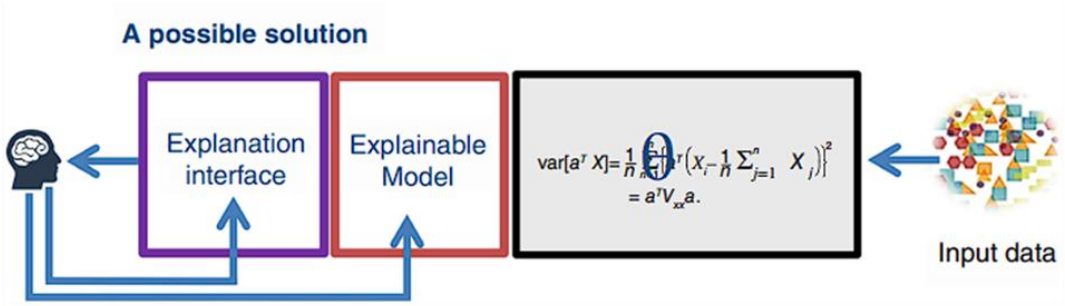

Figure 1 – A possible solution according to Holzinger et al. (2019)

Black-box models like deep neural networks, ensemble methods and others on the other hand are not transparent by its architecture, which is where XAI comes into play, as their internal decision logic has to be made understandable. They achieve high accuracy and can handle raw, high dimensional data which makes them opaque and thus a black-box.

They bring a risk when they are used in the medical field, because without proper explanations, clinicians cannot validate and trust its predictions. The best practical solution is, according to Holzinger et al. (2019) a hybrid solution which uses a black-box model for their predictive power and on top of that, explanations to surface human understandable reasoning with a focus on human causality (Fig. 1).

5. Explanation Techniques

Post hoc explanation methods are applied after the model was trained and aims to interpret the model without requiring changes to its architecture. Methods like saliency (and others) are used to derive sensitivity of the output to inputs (which pixels/features contributed). Local surrogate models like LIME approximate complex model behavior locally with interpretable models for a particular instance. These methods allow explanations without redesigning the model, giving insights into the model's decision and behavior over individual samples. However, the local explanations may misrepresent the model's global behavior and can be misleading if interpreted as causal reasoning.

Ante hoc explanation on the other hand means that the model is explainable by design, such as GAMs, decision trees and other rule-based systems. For these models the decision logic is visible to humans, they can see the effect of features and trace the path of decision. Because they are designed to be interpretable, it is easier to validate and deploy them in high stake settings as they do not require any post hoc translation. Interpretable models are limited in their flexibility and accuracy where complex non-linear representations are needed.

6. Example Use Case

Within Holzinger et al.'s paper, a histopathology use case (Fig. 2) was used as an example of the challenges of explainability in medicine. The authors asked an experienced pathologist to provide post hoc explanations through identifying relevant regions in histology slides and then ante hoc explanations, describing the structured reasoning process leading to diagnosis. This exercise highlighted the vast difference between human causal reasoning and AI's data-driven decision-making. Expert knowledge in histopathology relies on years of training and experience, integrating multiple features with varying weights and causal relationships to reach a diagnosis. The paper's inclusion of this example is very important, as it demonstrates the complexity of medical reasoning and makes the abstract concept of causability much easier to grasp. From the detailed ante hoc description, it becomes evident that replicating such nuanced human interpretation in AI systems remains a difficult challenge.

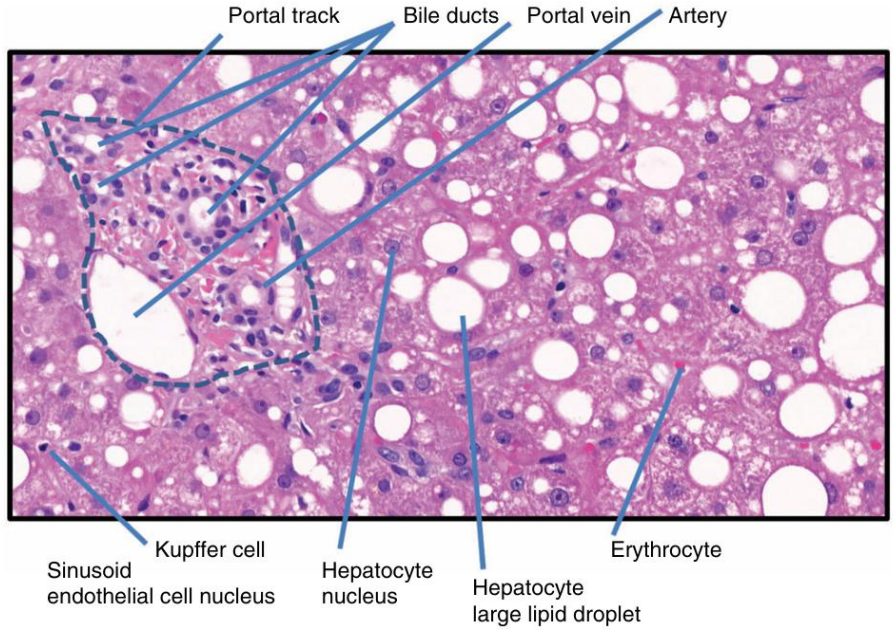

Figure 2: Features in a histology slide annotated by a human expert pathologist.

7. Future Work

In terms of future directions, Holzinger et al. discuss several promising and challenging areas. The first is weakly supervised learning, which could help overcome the expense and limitations of fully supervised learning. In medical image recognition, the datasets are often small, labeling is costly, and ground truth data are limited. The authors propose using weak supervision, where training data are incomplete, inexact, or inaccurate, combined with human observation during diagnosis to generate relevance maps and link known medical features with machine-learned patterns. Important to notice is that this approach could reduce training costs and provide valuable insights, it remains dependent on human expertise and is typically less accurate than fully supervised models.

Another proposed direction is the development of structural causal models. The authors argue that current AI operates mainly in statistical or model-free modes, which restricts its ability to reason about interventions and causality. The authors call for AI systems to incorporate models of reality, supported by visualization techniques that can be trained by medical experts. Developing structural causal models of human decision-making and integrating them into deep learning systems could, in theory, lead to more interpretable and trustworthy AI. However, this vision is highly ambitious and faces significant obstacles in our opinion such as data sparsity and the complexity of merging causal inference with deep learning. As there is no empirical validation that this could work, it should be viewed as an important research direction rather than a worked-out solution.

The authors further advocate for establishing causability as a new scientific field, with its own methodologies, metrics, and evaluation frameworks. They propose a “systems causability scale” to assess the quality of explanations. This would involve evaluating causal explanations in terms of their effectiveness, efficiency, and user satisfaction. In the medical context, the authors talk about how such a framework might address questions like: “How would seeing X change my belief in Y?”, “What if I do X?”, and “Was Y the cause of X?”. Although the concept is insightful, the proposed metrics remain subjective. Nevertheless, the idea of causability as a distinct performance measure, separate from explainability, represents an innovative direction for future AI research.

8. Conclusion

In conclusion, Holzinger et al. correctly emphasize that medical AI must enable doctors to understand how decisions are made and to assess the quality of explanations. The paper's conceptual contribution is strong, yet its framework remains largely theoretical. There is a notable absence of real-world empirical validation showing that causability improves clinical outcomes. Moreover, the distinction between explainability (as a system property) and causability (as a human property) may be overly simplistic, as in practice these dimensions are deeply intertwined, particularly in the design of explanation interfaces that shape human understanding. Future work should also explore AI systems capable of generating meaningful explanations without relying heavily on human experts.

Overall, this paper is a thought-provoking and forward-looking contribution. It serves as a conceptual manifesto that opens an important line of future work rather than offering a complete scientific framework. Its strength lies in articulating the necessity of moving from mere explainability to causability, encouraging future empirical research to operationalize and validate this idea in practical, clinical AI systems.

References

- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4). https://doi.org/10.1002/widm.1312