Interpretability of polypharmacy deep-learning models

Lynda Dib

Department of Computer Science and Software Engineering,
Université Laval, Québec, Canada
Lynda.dib.1@ulaval.ca

Abstract

Despite the successes of deep learning algorithms, they suffer from a fundamental issue called the "Black Box" (BB) effect. This effect is due to the lack of interpretability of the model and of explainability of the predicted results. The aim of my work is to propose solutions to this issue in the context of polypharmacy AI solutions.

1. Introduction

While deep learning (DL) models offer high-quality solutions in a wide range of problems, one major issue limits their adoption in critical areas such as medicine, insurance, and finance. These systems are "black boxes" (BB), meaning their internal workings are hard to understand and their decisions to explain. Combating this issue has given rise to the research area of eXplicable Artificial Intelligence (XAI) [1].

2. Background

Researchers have established a distinction between transparent models that can be interpreted by their design, and BB models which require external interpretability techniques to be understood [2–5]. A transparent AI system is one whose inner logic and decisions can be understood directly [2] by humans. This category includes regression algorithms, decision tree, K-nearest neighbors, rules-based learning, general additive models, or Bayesian models [6]. By contrast, the inner workings of DL models cannot be understood directly; in fact, a lot of the high performance of DL solutions results from their inner complexity. These BB systems require separate additional methods to analyze and explain their decisions [1, 7–11]. These solutions are often classified along two dimensions [5, 6, 12]. First they can be **model-agnostic**, meaning they offer a posterioiry interpretability of the results regardless of the AI technique used to generate them, or **model-specific**, meaning they are adapted to a specific BB model. And secondly, they can be **global**, meaning they explain the overall behaviour of the model, or **local**, meaning they explain a specific prediction.
In current papers, almost no work describes what an interpretable model should impose from a conceptual point of view. All existing interpretation models are developed for a specific purpose. Currently, there is an ongoing debate on which method should be used? [11, 14–16].

3. Proposed interpretation approach

Polypharmacy is defined as the long-term intake of more than five drugs by the same person [13]. This situation describes two-thirds of Canadian seniors. It can have negative consequences due to the accumulation of side effects and unforeseen interactions between drugs. Discovering and predicting which polypharmacy situations are harmful is a major data mining and AI challenge.

* lynda.dib.1@ulaval.ca

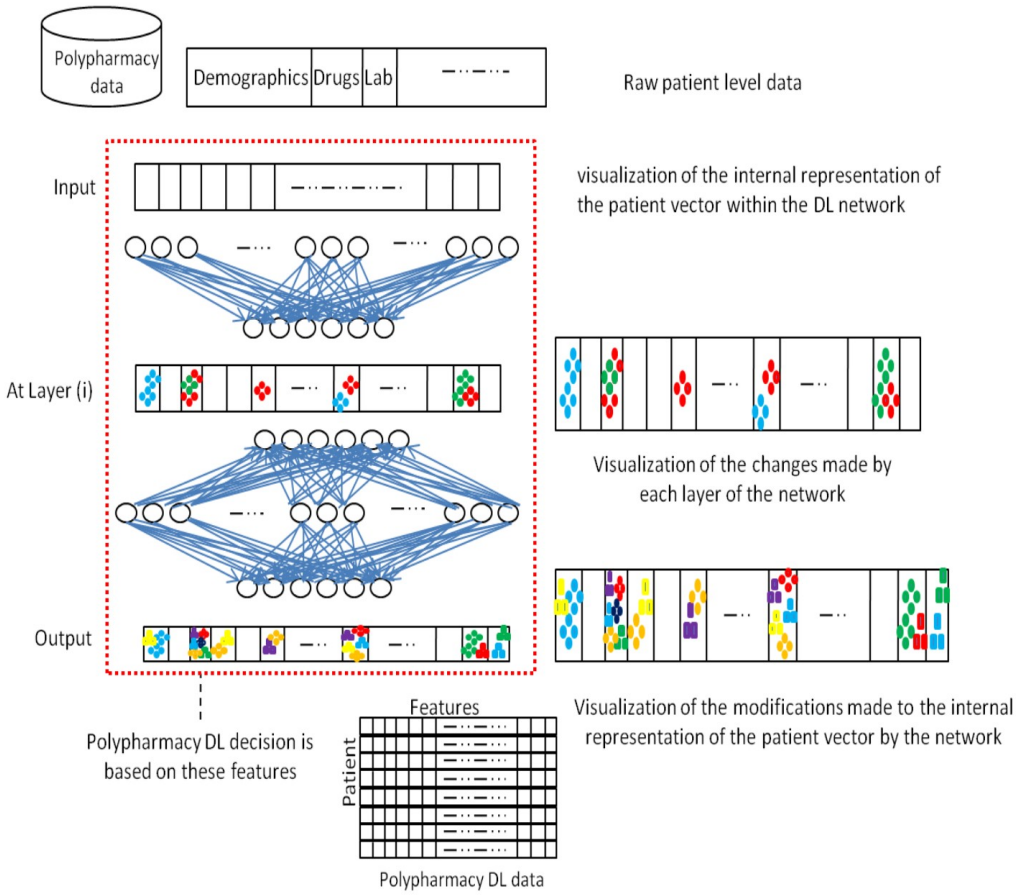

Figure 1. Visualisation of Patient vector modification in Polypharmacy DL model

My work is part of a research program that will notably lead to the development of a DL model to predict harmful polypharmacy. My research will focus on the interpretability of this polypharmacy DL model. Indeed, medical decisions must be explainable: it is not sufficient for an AI model to recommend adding or removing a drug from a patient's prescription, it must be able to explain to the medical professionals and the patients why these changes are needed. I propose to build an agnostic local model to visualize the reasoning behind these decisions.

More specifically, I will develop a visualization of the results generated by the internal mathematical operations performed by the network on the input patient data. This will allow us to better understand what is going on inside the model and how patient data affects drug recommendations. To do this, I will first develop a visualization of the internal representation of the patient vector within the DL network. I will then display on it the changes done to this representation by the network. This will include changes done by each layer of the network, how it connects to work done by the previous layers, and which component has undergone the most change. Expanding this to the entire network will make it possible to visualize which input values in the patient's vector have had the most impact on the output decision (see Figure 1).

4. Conclusion

Interpretability is a necessary complement to a highly-performing DL system, especially for problems such as drug prescription which demand explanations of decisions. A visualization approach is the best way to convey and explain complex information to health experts and laypeople, and thus promote confidence in the solution provided by the system. My propose approach will make it easier for all users to understand how input data is processed by the DL system and which parts of it most influence the output decision.

Acknowledgements

This research is funded by CIHR/NSERC collaborative grant number PRCS 549730 and supported by the INSPQ. I am grateful to my supervisor Richard Khoury.

References

[1] W. J. Murdoch, C. Singhb, K. Kumbiera, R. Abbasi-Aslb, and B. Yua. "Definitions, methods and applications in interpretable machine learning". In: *PNAS* 116.44 (1995), 22071—22080.

[2] A. Barredo Arrieta, N. Díaz-Rodríguez, and J. Del Ser. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Elsevier Journal of Information Fusion* 58 (2020), pp. 82–115.

[3] D. Finale and B. KIM. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. URL: https://arxiv.org/abs/1702.08608.

[4] B. T. Doran D. Schulz S. *What does explainable AI really mean? a new conceptualization of perspectives*. 2017. URL: https://arxiv.org/abs/1710.00794v1.

[5] A. Choudhary. *An Important Introduction to Interpretable Machine Learning Models in Python*. 2019. URL: HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/AUTHOR/ANKIT2106.

[6] C. Molnar. *Pitfalls to avoid when interpreting machine learning models*. 2020. URL: https://arxiv.org/abs/2007.04131.

[7] A. Alex. "Comprehensible classification models". In: *ACM SIGKDD E. NL.* 2014, pp. 1–10.

[8] R. Andrews, J. Diederich, and A. B. Tickle. "Survey and critique of techniques for extracting rules from trained artificial neural networks". In: *Knowledge-based systems* 8.6 (1995), pp. 373–389.

[9] U. Johansson, R. Konig, and L. Niklasson. "The truth is in there rule extraction from opaque models using genetic programming". In: *CAI.* 2004, 658—663.

[10] C. Meske and E. Bunde. *Using explainable artificial intelligence to increase trust in computer vision*. 2020. URL: https://dev.arxiv.org/abs/2002.01543v1.

[11] B. Boehmke and B. Greenwell. *Hands-on machine learning with R*. 2020. URL: https://bradleyboehmke.github.io/HOML.

[12] V. Buhrmester, D. Münch, and M. Arens. *Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey Analysis*. 2019. URL: https://arxiv.org/abs/1911.12116.

[13] N. Masnoon, S. Shakib, L. Kalisch-Ellett, and G. E. Caughey. "What is polypharmacy? A systematic review of definitions". In: *BMC Geriatr* 17.230 (2017), 3—10.