© 2022 European Federation for Medical Informatics (EFMI) and IOS Press.

This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

doi:10.3233/SHT1220407

## Machine Learning in Medicine: To Explain, or Not to Explain, That Is the Question

Cedric BOUSQUET<sup>a,1</sup> and Diva Beltramin<sup>b</sup>

<sup>a</sup>Sorbonne Université, Inserm, université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006 Paris, France b'Service de santé publique et information médicale, CHU de Saint Etienne, France

**Abstract.** In 2022, the Medical Informatics Europe conference created a special topic called "Challenges of trustable AI and added-value on health" which was centered around the theme of eXplainable Artificial Intelligence. Unfortunately, two opposite views remain for biomedical applications of machine learning: accepting to use reliable but opaque models, vs. enforce models to be explainable. In this contribution we discuss these two opposite approaches and illustrate with examples the differences between them.

**Keywords.** Artificial intelligence, Explainability, Algorithms, Decision making, Trust

## 1. Introduction

Artificial intelligence systems are not yet used significantly in practice, which raises the question of their perceived usefulness [1], trustworthiness [2] and acceptance [3] by physicians. In 2022, the Medical Informatics Europe conference created a special topic called "Challenges of trustable AI and added-value on health" which was centered around the theme of eXplainable Artificial Intelligence [4]. The main idea is that new AI systems operate like black boxes [5] and cannot explain their predictions. We agree that this raises problems for the acceptance of such systems by physicians. However, imposing explainability as a mandatory criterion could lead to not being able to use the most efficient algorithms in certain situations. After presenting arguments in favor and then against explainibility of black boxes, we show that opinions are less clear-cut than it seems at first glance.

## 2. Results and discussion

Some models present inherent explainability such as decision trees where the explanation is the model. In applications on medical imaging, genomic data or hospital reports, data and models are very complex. The relationship between inputs and outputs is not linear and cannot be explained by a simple model, which requires post-hoc interpretation [6]. One solution is to build, in parallel with the black box, a model for which we know the

<sup>&</sup>lt;sup>1</sup> Corresponding Author, Dr Cedric Bousquet, SSPIM, Bâtiment CIM42, chemin de la Marandière, Hôpital Nord, 42055 Saint Etienne; E-mail: cedric.bousquet@chu-st-etienne.fr.

execution logic, for example a linear regression model, by trying to make its response correspond as well as possible to the response of the black box. However, the substitution model presents lower performances than those of the black box. In the case of medical imaging, this is generally performed using saliency maps [7]. Post-hoc interpretations of black box predictions generate explanations that are not necessarily related to these predictions. This can lead to an erroneous belief that we understand the predictions of the model when we do not know its internal behavior [8].

Additionally, these explanation algorithms lack robustness because they may not generate similar explainations when confronted to similar data, as a small change in data may lead to different explanations. It is also difficult to retrace the steps when the system makes an error using post-hoc interpretation because it imperfectly approximates the function leading to the prediction. Saliency maps applied to images make it possible for the clinician to see where the regions of interest are, but do not indicate what is the characteristic that makes the regions interesting for the system [6]. Moreover, Abedayo et al. showed that the ability to evaluate predictions through a visual representation can mislead the user [9]. Experimentally, they established that some salience methods are independent of the model and the way the data is generated. Explainability is primarily intended for the system's implementers [6] who can thus determine why their system is not providing the correct answer, and thus improve it.

In some cases, it is truly impossible to produce an explanation that makes sense to the user. An example of this is DeepGestalt, an algorithm capable of recognizing facial phenotypes associated with genetic disorders; the system works well, however it is unable to give explanations to the end user of what were the features that induced the final diagnosis [10]. Explainability is certainly desirable and, although we hope that it will be implemented in future algorithms by design, it should not be constrained to the detriment of the use of AI systems that are difficult to explain but which can be of service to the patient.

## References

- Hah H, Goldin DS. How Clinicians Perceive Artificial Intelligence-Assisted Technologies in Diagnostic Decision Making: Mixed Methods Approach. J Med Internet Res. 2021;23(12):e33540. doi: 10.2196/33540.
- [2] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655. doi: 10.1016/j.jbi.2020.103655.
- [3] Bates DW, Kuperman GJ, Wang S et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc. 2003;10(6):523-30.
- [4] Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. Front Big Data. 2021;4:688969. doi: 10.3389/fdata.2021.688969. eCollection 2021.
- [5] Kundu S. AI in medicine must be explainable. Nat Med. 2021;27(8):1328.
- [6] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750. doi: 10.1016/S2589-7500(21)00208-9.
- [7] Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med. 2021;140:105111.
- [8] Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284-286.
- [9] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. arXiv preprint 2018. arXiv:1810.03292.
- [10] Gurovich Y, Hanani Y, Bar O et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25(1):60-64.