Trustworthy Artificial Intelligence in Medical Applications: A Mini Survey

Mohsen Abbaspour Onari[1], Isel Grau[1], Marco S. Nobile[2], and Yingqian Zhang[1]

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
{m.abbaspour.onari, i.d.c.grau.garcia, yqzhang@tue.nl
[2] Ca' Foscari University of Venice, Venice, Italy
marco.nobile@unive.it

Publication: The full paper of this abstract has been accepted at CIBCB '23. Available at: https://ieeexplore.ieee.org/abstract/document/10264883

1 Introduction

Data is being generated in many fields on a large scale which offers the opportunity to use Artificial Intelligence (AI) and Machine Learning (ML) models to automatize decision-making for many domains, including healthcare, education, cybersecurity, manufacturing, and transportation. However, there are equally dire concerns regarding biased decision-making, unfair treatment of minority groups, privacy violations, adversarial attacks, and challenges to human rights. The leading cause of this concern is that most ML algorithms produce a complex model that makes them black boxes. Besides, the immense amount of data might include human biases and prejudices. Consequently, decision models learned from them might inherit such biases, possibly leading to unfair and wrong decisions. To maximize the benefits of AI and mitigate or even prevent its risks and dangers simultaneously, the Independent High-Level Expert Group on Artificial Intelligence has provided a guideline to obtain trustworthy AI.

2 Research importance

In this paper, we focus on algorithmic requirements and limit our study to analyze their real fulfillment in practice for five requirements: accuracy, transparency, trust, robustness, and fairness. In this review paper, we analyze the fulfillment of these algorithmic requirements to establish trustworthy AI through the lens of the healthcare and medical field literature. The primary motivation to do so is that medical experts might not understand the prediction process, which leads to a relatively low acceptance of AI. Moreover, it could limit the possibility of complementing evidence-based medicine with ML models for diagnosis and treatment decisions. Our results will show that in addition to that, there is a long way to achieve the fulfillment of trustworthy AI in practice. Most studies consider explainability as the ending point of the research and do not fulfill the remaining requirements. It is worth mentioning that based on the domain experts' expectations and research goals, the importance of the algorithmic requirements may vary. It turns out that there must be a trade-off between the algorithmic requirements of trustworthy AI based on the application domain. Finally, we will point to developing multi-objective optimization problems as a potential solution to fulfilling trustworthy AI. In this study, we do not go through the modeling of such a problem, and we only want to attract researchers' attention to evaluate the possibility of using this solution to develop a trustworthy AI decision support model.

3 Results

In this mini-survey, we studied the fulfillment of algorithmic requirements to establish trustworthy AI in the medical domain. The primary motivation is the lack of tendency to rely blindly on AI systems by medical experts due to the black-box nature of advanced ML models. Our approach is based on studying XAI methods that can explain the black box model's decision. However, because XAI is not the only requirement to establish trustworthy AI, we considered five algorithmic requirements: accuracy, transparency, trust, robustness, and fairness. In order to track the studies easily, we separated the covered research papers into general ML and DL models. Our findings show that most authors consider providing explanations as the ending point of the research, which is not a correct approach. The validity and quality of the provided explanations are not evaluated through experts' knowledge, which is a big concern. Next, most research papers poorly declared that the medical experts' trust is obtained by providing explainability. Implementing XAI does not guarantee to achieve trust in practice, and the trust should be quantified by a methodology to model the experts' mental model. For robustness, it is observed that researchers only consider some techniques, such as cross-validation, feature engineering, and tackling missing values, to have a robust prediction. However, based on trustworthy AI requirements, the robustness of explanations and the resilience of the AI system in the presence of errors and outside attacks must also be considered. For the fairness requirement, researchers do not address the accountability and responsibility concerns facing an unethical decision. Besides, the impact of explanations to track causal steps ending in the prediction needs to be included. The results demonstrate that there is still a long way to achieve trustworthy AI in the medical domain, and researchers need to consider all algorithmic requirements in the future. As a potential solution, we want to attract researchers' attention to develop multi-objective optimization problems to make a trade-off between different algorithmic requirements based on the problem and domain experts' expectations. This survey is the beginning step for a more extensive study to provide a bigger picture of fulfilling algorithmic requirements in trustworthy AI. We restricted our analysis to the Scopus database to cover research papers. For future studies, we will go through the other databases, including Google Scholar, PubMed, and Web of Science, to comprehensively review this domain.