Review

Causability and explainability of artificial intelligence in medicine: A Review

By Aymen Bashir

Andreas Holzinger1 | Georg Langs2 | Helmut Denk3 | Kurt Zatloukal3 | Heimo Müller1,3

Introduction

This research paper presents an in-depth discussion on the growing Human-AI relationship in medicine. There is no denying that AI is performing exceptionally well in the areas such as autonomous driving, speech recognition, face detection, and recommendation system due to huge data sets and commendable computation power. Yet, the AI results have certain limitations concerning effectiveness and performance when it comes to interventions and retrospections. Thus, applying the deep learning solutions to medical problems gets under debate due to a lack of explanation of artificial neural networks methodology. The idea of probabilistic learning enhanced the success rate but, at the same time, made the process increasingly opaque too. Thus, to tackle these problems, the concept of explainable AI has been brought to notice in this research paper. The idea of explainable AI tends to bring out traceability and transparency regarding black-box machine learning methodologies meant for medical solutions.

This paper discusses in depth the following two terms:

- **Explainability:** This does not explicitly refer to a human-understandable model. Instead, it tends to highlight the parts of the algorithm that are relevant to the decision and actively contribute to the model's accuracy.
- **Causability:** It is measured in terms of the effectiveness, efficiency, and satisfaction related to the causal understanding and its transparent relationship with the user. In short, it refers to such a model that a human can understand.

While building upon these very terms, the idea of explainable medicine would ultimately require explainable medicine. Moreover, this paper also differentiates between explainability and causablity with the help of use-case in histopathology. Finally, the notion of causability gets explained in terms of the property of a person, while explainability is the property of a system.

Critical Analysis:

This paper tends to sort out the underlying methodology that works the Blackbox architecture of an artificial neural network. The major strength of this paper is that it integrates the importance of explainable models in medicine. This paper helps us understand the fact that explainable AI is not only necessary and useful but also presents us with a huge opportunity when it comes to fields such as medicine. The importance of causability and explainability in the medical domain is specifically necessary because there is no well-defined ground truth for medical diagnosis. Medical, scientific models are also based upon causality as a final goal for understanding the underlying mechanism. Thus, this paper suggests making AI models more explainable and causable as such models would understand the context. Therefore, one could reason about the intervention and retrospection as these factors are recursive for the medical field.

This paper discusses various models and techniques that ensure explainability and causability, such as activation maximization, attribution, posthoc and ante-hoc systems. Although these techniques help us know the underlying process yet they do not perform exceptionally well. The most successful models currently used in machine learning are statistical or model-free modes. These models can't help us with interpreting the context, and we are not also not entirely familiar with their work, yet they give us SOTA results. This paper fails to highlight tools and techniques that should be incorporated with explainable models to outperform these current models.

The most significant aspect this paper presents is the idea of explainability. This very step could help us bridge the human-AI gap. In the field of medicine, one has to adapt and improvise in various scenarios. For that, intellectual understanding is necessary. Moreover, if one employs opaque deep learning models for SOTA results, one would require a huge chunk of labeled and clean data, which is a very expensive medical method and results from the process. Thus, this paper promotes unsupervised learning in medicine with explainable models based on posthoc and ante-hoc systems (GAMs). Moreover, to understand and explain the models, this paper features attribution and activation maximization techniques that map the input and output relationships via the proper steps involved. Finally, this paper hints to develop causability as a new scientific domain as this helps explain the scientific theories in AI. This would help us with the quality of explains in ANNs.

One of the major advantages of explainable and causable models is that once we know the underlying mechanism a machine follows, a domain expert, in our case a medical professional, could link his classical theories with the machine's approach. Once we can derive a correlation between the two said approaches, we could understand the machine better. In turn, we could feed the ANN with relevant data with the information it feels necessary instead of applying brute force to make it train over a huge chunk of data over an extended period.

When it comes to deploying and using the explainable models, this paper mentions three different explanations. Among them, this paper targets the peer-to-peer description as it is carried out between a group of physicians during medical reporting. Thus, this paper carries significant managerial significance when it comes to medicine. It helps to bridge the gap and develop trust in AI-based medical professionals' results as the model's approach could now be perceived and understood accordingly.

The paper under review has featured an extensive literature review. After thoroughly researching a significant amount of background information, the authors have discussed the importance of the medical field's said technique. After making a base on substantial background data, the authors have made an adequate analysis of the importance of artificial intelligence's causability and explainability in the medical field. The authors have explained the relation of explainable models to the medical field and the methodology and steps these models follow to make their working transparent and understandable.

This paper contributes to explainable AI with the help of a use case of histopathology in the medical domain. Initially, the authors have explained the importance of explainable models in the medical field as it lacks well-defined ground truths. The data is scarce, poorly labeled, and expensive to acquire, and medical professionals must find a way to trust the machine's decisions. Thus, this paper could serve as a huge stepping stone as it employs explainable AI to bridge the gap between medical professionals and ANNs.

This paper was a well-organized piece of work. However, the visual aid could have been more descriptive as it was a vital part of the discussion under consideration.

Overall, the authors have clearly explained the problem statement and have supported their approach with the help of a case relevant to the field. A future outlook has also been added that focuses on weakly supervised learning, structural causal models, and an approach to develop causability as a new scientific field. However, the part that describes the structure and architecture of explainable models could have been made more elaborate and briefer. Furthermore, the paper should have also had more comparative analysis for solvable models and models that give SOTA results in the medical field but are opaque.

This paper has driven huge stress over the terms of explainability and causability. These approaches were discussed so that they could be employed to overcome the practical challenges in medicine.

The major emphasis area has been that AI models must be understood instead of brutely feeding them with data. As if an expert could understand the way they work, they could derive a relationship between the underlying theories that could help them optimize the model better by feeding it relevant data and saving valuable time and cost incurred for acquiring data.

References:

1. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608.
2. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349, 255–260.
3. Montavon, G., Samek, W., & Müller, K.-R. (2017). Methods for interpreting and understanding deep neural networks. arXiv:1706.07979.
4. Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, 23, 351–379.
5. Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331, 1279–1285.