Critical Review:

Causability and explainability of artificial intelligence in medicine

In the field of medicine, there is a growing demand for AI models that not only perform well but are understandable, explainable, and interpretable. In this paper, Holzinger highlights the concept of explainability and causability and how they assist in comprehending an AI model. Explainability interprets the decision-relevant parts of the model whereas causability is the extent of human understanding while looking at an explanation. Throughout the paper, the authors discuss the need and ways of reducing the opacity in DL/AI/ML through various explainable AI models. Post-hoc and ante-hoc explainable systems, their differences, standards, and shortcomings were evaluated in detail. Furthermore, it is suggested to reinforce explainable models with causability. The author also confers about various methods that aid in interpreting a deep neural network i.e. uncertainty, attribution, and activation maximization. To practically demonstrate explainability with causability, the authors have provided an example where a professional provides insights on liver pathology through post-hoc and ante-hoc explanations. Given the expense and uncertainty surrounding supervised learning and the imbalanced datasets, the author proposes to research weakly supervised learning algorithms, establish causability as a new field and build structural causal models to ensure AI achieves human-level intelligence.The authors of the paper have well researched the area and several citations are provided. This paper briefly states the need of explainability and causability in the AI medicine. Intensive background information is provided and existing AI models are discussed. Shortcomings and issues related to data, models (Deep learning, explainable AI) are also highlighted concisely. Assumptions while modelling an approach were mathematically expressed. Example based on the explanations of a professional using post-hoc and ante-hoc explainable model provided the reader with in-depth knowledge of the approaches.

The main notion addressed in this paper is causability. Explainability assists in providing professionals with insight on “how” the prediction was made whereas causability will assist them in understanding the “why”. The proposed method is feasible and practical (as suggested by the example). This approach turns the black box into a white glass where everything is transparent. In the medical domain, augmenting explainability and causability will enhance the trust in AI, therefore, resulting in better diagnosis.

Undoubtedly, exploring areas like weak supervised learning and causability will build human-AI trust however, many other concerns are raised simultaneously. In the medical domain, acquired data is quite imbalanced, missing, and inaccurate hence needs effective preprocessing as AI is entirely data-dependent. Furthermore, causability is dependent on a person’s perspective and it may vary from person to person and might lead to inaccurate models.

Overall, the paper is well-explained however, formatting is required. Mathematical equations and notations need to be labeled properly. The paper lacks in explaining existing approaches of explainable AI in depth. Understanding can be enhanced by adding visualizations and examples. The author has discussed miscellaneous interpretation techniques within one section, therefore, resulting in a lack of clarity. The addition of sub-headings will give a formal look and make the text more understandable. More examples and

research work is to be cited to support the discussed approaches. Moreover, the paper doesn't provide any performance evaluation to analyze the approach realistically. Lastly, the papers cited are quite old therefore authors need to refer to the latest research in the domain.

In conclusion, the implementation of causability together with explainable AI will lead to a dramatic improvement in the human-AI relationship with respect to medicine. The research paper provides knowledge regarding the existing AI models however the research is quite biased as the paper emphasizes the issues regarding causability. Practical examples are to be considered for future research. The paper acts as a basis for any research in the medical domain in terms of explainability and causability.