Critical Review

Causability and explainability of artificial intelligence in medicine

This paper by Holzinger et al discusses the contributions of Artificial Intelligence in the field of medicine making it more transparent to a user. The author discusses the integration of explainable AI, Deep learning and Machine learning in medicine and the approaches are now making the process more robust while after that it talks about the importance of knowing the causality of a decision. It presses upon the necessity of medicine being a sensitive concern as it is directly linked with the lives of patients, therefore the need of explainability followed by causality is the major topic in discussion. Moreover, the difference between causabilty and explainability is highlighted by the user, explainability being a property of an AI system while causability being the property of users. Therefore, a causable explainable AI system will help users understand the decision making process by interfacing transparency into the AI system.

Furthermore, the author discusses two types of explainable AI systems that are: (1) Post-hoc Systems and (2) Ante-hoc systems. He mentions the ground differences between the two approaches as the post-hoc system has a model-agnostic approach also known as black box approach and on the contrary ante-hoc systems have an interpretable nature following a glass box approach.

Moreover, the authors revisit general advancements in explainable AI, with an emphasis on deep learning and the approaches being made to explain neural networks. The techniques include uncertainty, attribution and activation maximization. It further explains the usage of unsupervised learning and its benefits on the formation of the algorithm on the basis of input data.The paper also mentions examples of both explainable AI systems and analysis done under the supervision of a professional.

Lastly, it mentions the future enhancements of the combined approaches that include: (1) Weakly Supervised Learning due to minimum availability of labeled data, (2) Structural causal models to achieve human level intelligence and (3) development of causability as a new scientific field to ensure the quality of explanations being reproduces.

The respective paper is a review of advancements of Explainable AI with its accompanying methodologies, weaknesses and requirements. The paper is supporting its claims from references making it an authentic resource for discussion. It entirely covers the ideologies being used such as Deep Learning and the work being done to understand the working behind it. The main motivation of integrating causable explainable AI systems is well

explained by the author and its increased necessity and domains have also been highlighted adequately. Moreover, the paper identifies the difference between the two terms: *causability* and *explainability* and emphasizes the necessity of both of them. This paper ensures the reader to understand the importance of a glass-box based approach in AI to make the decision process more robust and trustable. Overall, the paper is able to convey its purpose ,backed up by references and examples.

However, with the in-depth detail being provided by the authors, the paper is not well structured.There are no visualizations or representations of practical implementation of the proposed methods.Details have been flooded in different headings that are unnecessary keeping in view the main motivation of the authors. There has been no practical approach to different approaches it has explained e.g in the case of understanding deep neural networks, no examples have been cited for different techniques mentioned there. There is vivid shifts of topics inside the same heading making it hard to analyze the main objective of that specific paragraph

Furthermore, the paper focuses on the importance of causability and explainability which endangers data privacy and security of the patient, being a big hurdle in the pathway. Lastly, the authors do not mention any method of analyzing the performance of an Causable Explainable AI, as causability is intuitive in nature and can only be cross checked by manual means. Therefore, the paper does not highlight the limitations that are adjoined with the following approaches.

In conclusion, the following paper has been able to convey the main motivation i.e integrating causability in Explainable AI models designed for medicinal use. The details and in depth knowledge presented by the authors, indicate a great amount of research being done for it. It sums up the future requirements of AI in the medicine field and mentions the future developments required for AI to be trustable. The following limitations mentioned above include practical approaches and performance analysis methods for it so that analyzing the progress becomes possible.