**Causability and Explainability of Artificial Intelligence in Medicine**

**Paper Review**

**Summary:**

Artificial intelligence is one of the oldest fields of computer science. The authors raised the point that with the progress of statistical models, large data and cheaper computational resources has helped to produce results that are never seen before however it has made the process more opaque. The authors focus on the point that even though ML models are working great in the field of medicine but this field is an applicable field with dangers to human life involved so the need of a human expert is always needed. This can only be possible if doctors understand how and why a machine has made the decision.

The authors move forward with differentiating between causability and explainability. Explainability is part of the algorithmic model that is used while predicting an outcome. Causability is however the extent of understanding a human gain while looking at a statement.

The author also mentions two systems post-hoc and ante-hoc. Post-hoc systems are the ones which produce the explanation after the prediction has been made thus work as a wrapper on the black box model and also called model-agnostic system. Ante-hoc systems are interpretable by design towards glass-box approaches. These types of the systems are required in the medical field as experts can combine their knowledge with the hidden knowledge from the data. Recently the issues with scalability and accuracy have been solved so these models work in bioinformatics application.

Several approaches are made to explain Deep learning neural method. The techniques involve uncertainty, attribution, activation maximation etc. The authors also show how humans ask questions in the post hoc and ante hoc systems.

The future improvements the author propose includes using of weakly supervised models in the medical domain as labeled data is not in excess and not in the form to be used by our models. Secondly to have research towards structural causal models so AI can achieve human level intelligence. They also propose to establish a field of causability that looks in quality of explanation as it will be necessary with the progress of AI as an important technology.

**Critical Analysis:**

This paper is in line with the needs of medical field. AI has been playing major role in few fields and its time that medicine also benefits. The reason of this paper seems to focus on type of AI models that are understandable by human experts. Also, they coin the new term of ‘causability’ which is the extent to which one understands the result.

The strength of this paper is surely in how it explains the need of explainable AI in medicine field. It uses definition to make us understand what is the different interpretation of understandability means. That there is a mistrust between medical community and machine learning models. So, this paper explains the need of glass-based approach in AI models rather than black box models as medicine is the field that literally saves human lives and they should be in foremost list to benefit from advancement of AI. The use case of histopathology was a great example included in the paper and was also quite extensive and allows the reader to see how the two systems differ.

The weakness of this paper is not successfully explaining the present approaches of explainable AI already in use. There was a mention of a framework for visually explaining the decisions of any classifier and its use in bioinformatics application. The authors should have explained it more so the reader can fully grasp the knowledge of AI that is already present in the field of AI. There was a section which explains different techniques to interpret the readability of deep learning neural nets. The author could have used sub headings to explain the techniques as they mix up in the same section. Also, the authors could have presented examples with these techniques and also mention in which field they are being used. Also, in the approaches section of the paper the authors seem to go off scope as the mentioned approaches were relevant but their details were not relevant to the topic. There was also no section of related work which seemed like not much research work that could have allowed the reader to further understand what the authors were trying to convey.

**Conclusion:**

This is a research-based paper as no new approach was proposed and was solely to introduce the need of explainable AI in the field of medicine and how current models are not being used because of their statical approach to the data. So, this may not be cited regularly in the future papers of explainable AI but referenced from time to time to show the need of the such AI. However, the research seemed biased and not presented the holistic view of the problem. It does not fully go over all the aspects of medical related AI. Overall this is a great paper to understand what to expect from the future AI in medicine but does not show a clear picture on how to achieve it.