CS 598: Project Proposal
Explaining a Machine Learning Decision to Physicians via Counterfactuals

Ammara Ashraf – aashra5
Zubair Lalani – zubairl2

Problem Statement

Physicians often face a significant barrier to trusting AI-driven tools due to the lack of transparency in how these models reach their conclusions. While AI models can offer valuable insights, the lack of reasoning displayed makes it difficult for physicians to understand and interpret the rationale behind the recommendations. This often fosters a disconnect, leading to hesitation in incorporating AI tools into clinical decision-making.

To address this issue, the authors of the paper “Explaining a Machine Learning Decision to Physicians via Counterfactuals”, propose an approach that explains the AI model's decision-making process by leveraging counterfactual explanations. Counterfactuals offer actionable insights by suggesting what would need to change in a patient's data for a different decision to be made. For example, rather than simply telling a physician that a specific medication is not suitable for a patient, the model would provide a counterfactual explanation such as, “This medication would be appropriate if the patient's cholesterol level were 10% lower.” This approach fosters trust among physicians by enhancing the transparency of the model’s decision-making, delivering clinically relevant insights, and offering clear, actionable guidance on potential interventions.

Specific Approach

The authors propose an approach to enhance trust between physicians and AI models by combining transparent decision-making with counterfactual explanations. The model underlying this approach is built on a black-box machine learning model, such as random forests or gradient-boosted trees, used for prediction tasks. In a black-box model, the reasoning behind the model’s conclusions is not easily interpretable due to the lack of reasoning provided. To address this issue, the authors implement a counterfactual explanation technique that generates alternative scenarios to show how slight changes in the input data could result in a different decision.

The counterfactual generation is formulated as an optimization problem, where the model searches for the minimal changes needed in the input features to alter the model’s prediction. This not only makes the decision-making process more transparent but also provides actionable insights for the physician.

The authors evaluate their approach using both qualitative and quantitative metrics. The qualitative evaluation involves physician feedback, assessing how easily they can understand and trust the model's counterfactual explanations. The quantitative evaluation includes metrics like accuracy, fidelity, and explanation quality to assess how well the counterfactuals align with human judgment and how actionable they are. Finally, the model’s performance is compared against baseline models to evaluate whether the counterfactual approach improves interpretability and trust in AI recommendations.

Novelty, Relevance, & Hypothesis of Approach

The approach provided in this paper presents a novel and clinically grounded approach to machine learning interpretability. It achieves this level of interpretability in healthcare by leveraging counterfactual explanations, as described previously. Unlike traditional methods, such as feature importance or attention mechanisms, this method provides actionable, patient-specific recommendations that clinicians can utilize to guide their diagnosis or treatment. Machine learning in clinical settings is often neglected due to the critical nature of the domain, as well as its lack of interpretability by clinicians. A key innovation by the authors is ensuring that the counterfactuals are minimal and medically plausible, addressing a critical gap in the practicality of ML explanations for real-world clinical settings.

The proposed optimization-based technique produces the smallest necessary changes to patient features that result in a flip of the model's prediction, which ensures both interpretability and feasibility. Compared to baseline methods that may suggest unrealistic or vague interventions, this approach provides targeted, understandable guidance, thus fostering greater trust and usability among physicians. The authors hypothesize that such explanations will be more interpretable, actionable, and ultimately more effective at supporting clinical decision-making, making their method a significant advancement in the field of explainable AI for medicine.

Ablations / Extensions Planned

There are a variety of ablations/extensions that we will attempt to implement during this project. It is difficult to determine which of these ablations/extensions are feasible due to time constraints; however, we will aim to complete at least one of the following extensions/ablations. One possible extension would be to add a physician feedback loop where we would add a mechanism for physicians to provide feedback on counterfactual decisions to refine future explanations. Another possible extension would be to integrate user-provided clinical constraints such as only including FDA-approved drugs. In addition, we could apply the proposed counterfactual method to other baseline models aside from the black-box models mentioned. Our final option that may be more complex is to extend the paper to support multi-classification problems rather than only focusing on binary classification problems.

Data Access and Implementation Details

The paper uses the MIMIC-III dataset, which is publicly available through the PhysioNet repository. Thus, our group will be able to receive access by following the instructions provided by the instructors of the course. These steps include completing the required CITI "Data or Specimens Only Research" training and requesting access through PhysioNet. Once approved, we will be able to download all the data necessary to reproduce the paper.

Once we obtain the data, there is a data preprocessing pipeline that is described by *Wang et al. (2020)*, which transforms the raw MIMIC-III data into hourly time series suitable for their model. This step is crucial for aligning the data format with the one used in the paper.

The authors have also provided their open-source implementation of the counterfactual method they proposed. It is available under the CFVAE repository on GitHub:
https://github.com/supriyanagesh94/CFVAE

The repository contains the codebase required to train and evaluate their counterfactual model. Our group will be using this code for our work on this project. As of now, the computation looks manageable to recreate given the time and resources, but we expect to uncover more challenges as we dig deeper into the study.

References

Nagesh, Anjali, and Xinya Chen. "Explaining a Machine Learning Decision to Physicians via Counterfactuals." **Proceedings of Machine Learning Research,** vol. 209, 2023, pp. 1-13. https://proceedings.mlr.press/v209/nagesh23a/nagesh23a.pdf.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In Proceedings of the ACM Conference on Health, Inference, and Learning, pages 222–235, 2020.