ML1030 – Capstone Project
Project Proposal

Submitted by: Michael Vasiliou

Executive Summary

The application of machine learning in a clinical diagnostic setting opens boundless opportunities to improve patient care and health outcomes. The adoption of machine learning in this setting is often hampered by the lack of interpretability of results that occurs in some machine learning models. Ensuring and improving interpretability of results in a clinical setting would improve the ability of clinicians to understand the output of the machine learning model, and as a result improve the uptake of this supportive technology in their practice.

Project background

The overall research project goal is to improve the ability and acceptance of Machine Learning in clinical decision making. The research project is utilizing multiple datasets including the IMDB movie review, and the MIMIC 3 clinical dataset. Through analyzing the disparate datasets, including the health care records and discharge summaries of 40,000 patients and comparing results and interpretability across multiple machine learning models the project aims to improve clinician trust, model interpretability, and uptake within a clinical setting

> *The output of the proposed system is an interpretable Knowledge Base, which can link the pattern groups, discovered characteristics of records, and patients' records together to shows "what" (disease), "who/where" (tracking patient records back) and "why" (discovered patterns) to interpret clinical notes for better clinical decision making.*
> - *Interpretability on Clinical Analysis from Pattern Disentanglement Insight (Appendix)*

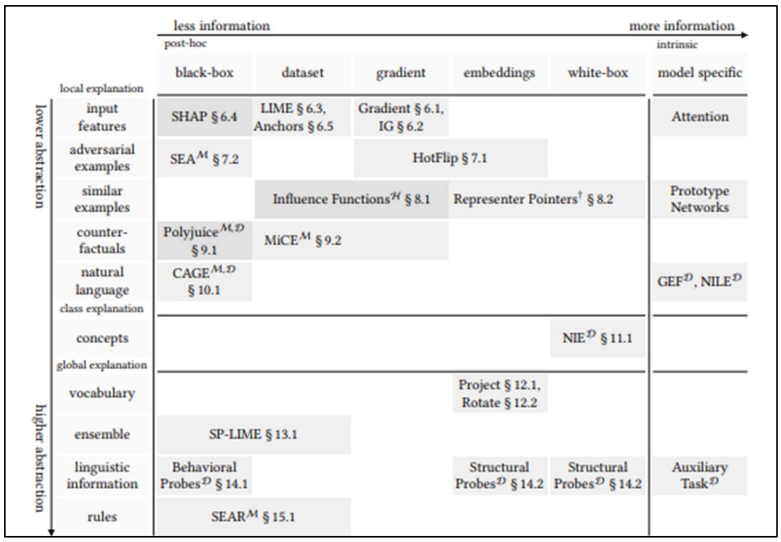
The main goal of the sub-project I will be focusing on is model interpretability. The chart below taken from "Post-hoc Interpretability for Neural NLP: A Survey by Madsen, Reddy, Chandar", shows a broad range of post-hoc interpretability options. Each interpretability choice has both strengths and weaknesses, and the desired project goals are the implementation of a broad range of interpretability options against the trained models, beginning with LIME, SHAP, Transformers-Interpret, and Concepts (NIE) and Vocabulary Projection and rotation.

Commented [1]: There should be at least 2 datasets
1. MIMIC 3 clinical dataset
2. IMDB movie review
3. other datasets the summer students will be collecting starting May 1st

Commented [2R1]: Added in a line for multiple datasets in the research and project.

Commented [3]: Can we say "beginning with LIME, SHAP, Transformers-Interpret" and "Concepts (NIE) and Vocabulary Projection and rotations"?

Commented [4R3]: Good idea. Changed as per your suggestion.



Implementation Strategy

The overall research project is a large project and has multiple moving pieces. To ensure success, several critical factors will need to be accounted for when designing and implementing a model interpretability and comparison solution.

1. Overall project will continue beyond the timeframe of this sub-project
2. Time constraints will limit the number of interpretability frameworks implemented
   - Initial interpretability methodologies to be implemented: LIME, SHAP, Transformers-Interpret, NIE (Natural Indirect Effect)
3. In the future, more resources will be added to the project to implement further interpretability options
4. Access to clinical dataset (MIMIC-III) being used is subject to privacy training and approval.
   - Privacy training was completed April 12th, but access approval to the dataset could still be several weeks out.

Interpretability Framework

The overall project already has both existing cleaned and completed datasets as well as multiple trained models: Random Forest, CNN, KMeans, PDD. With minimal coding intervention by users, the interpretability framework will need to support the following features:

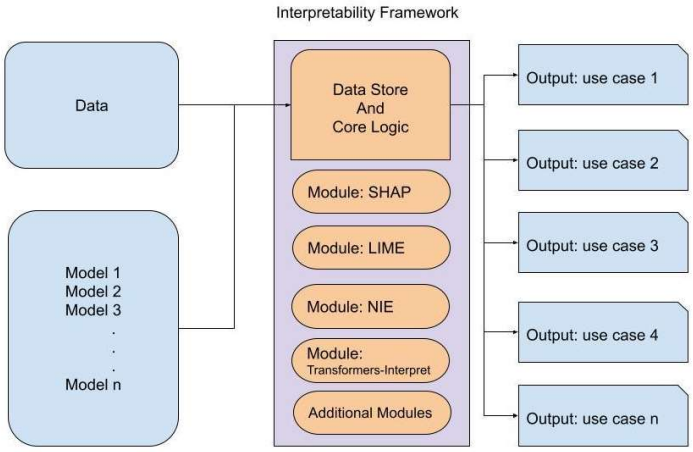
1) Load multiple models
2) Load model dataset(s)
3) Add interpretability frameworks (e.g., SHAP, LIME, NIE)

4) Output use cases

Use case options are not currently defined and will evolve as the project continues and more stakeholder feedback is acquired. Options for use cases could include different chart types, model to model comparisons at global level, and record to record comparison between models.

Commented [5]: Can we also spend some time on evaluating the outputs and comparing them to PDD? This is probably the more research portion of the work. You may consult Sean/Peiyuan on who might be assigned to work on this.

Commented [6R5]: Absolutely. Use cases should not be limited to single model/single output. Comparing multiple models concurrently seems like a critical function.



Project Timeline

The project will be divided into three main phases, the first of which is to understand the requirements and project goals, second is to create and deploy a minimal framework to project team members, and the third is to increase the breadth of use-cases, and interpretability frameworks.

Commented [7]: This is perfect! I was very confused up until now because in previous CSML1010 offerings this was done in the first class. Good to know that this has been moved to later; this is great because the team now a clear timeline and expectations.

Commented [8R7]: Thank you

Phase 1: [26-Mar-2022 16-Apr-2022]

- Requirements gathering
- Development environment configuration
- Privacy Training
- Proposal

Phase 2: [17-Apr-2022 29-Apr-2022]
Create and deploy to project team Interpretability Framework with key goals:

- Establish base framework and integration with project team
- Ensure code portability across development environments
- Base framework run on one project model for one type of interpretability chart
- Base framework run on all models for one type of interpretability chart
- Collect and document use cases

Phase 3: [30-Apr-2022 21-May-2022]

Development through Agile with frequent small deployments and release to the team after each new feature has been added.

- Based on project team priorities, iterate through:
  - Add frameworks
  - Add use cases
- Work with other team members to add/integrate new interpretability options

Risk Management

As this is a research project and not a production implementation there is a smaller but important set of risks we need to ensure are addressed.

Data Privacy

Risk: Clinical data is sensitive and requires privacy training and external approval prior to access. Data has been de-identified, so risk of privacy breach is not a concern.

Mitigation: Privacy training has been completed. When access has been approved, the clinical data will be stored in password encrypted, secured storage with no external access.

Timelines

Risk: The clinical data set needs external approval prior to being granted access. This will delay the ability to code the interpretability framework directly against the final dataset and model. Each model may require different code-paths when implementing and without coding directly against the final model/data there could be integration concerns

Mitigation: Working with other project team members, small iterations will be completed and deployed to team members to evaluate against clinical data and trained models. Frequent iterations should find any integration issues so they may be addressed early in the development. This will also have the benefit of beginning the integration and acceptance by the project team.

Project Continuity

Risk: As the research project has long term goals beyond the scope of this project, the interpretability framework must remain usable and maintainable by the project team

Mitigation: Throughout implementation, knowledge transfer, training, code-reviews, and inclusion of other project team members will be a priority. Implementation of additional interpretability modules will be documented for repeatability to add new functionality. A target goal is to have another team member implement a new interpretability module and have them walk through integration to the larger framework.

All code and documentation will reside with the full project team under their GitHub repository.

Commented [9]: It would be great if you could prepare a quick tutorial about "Agile method" and share it with the team and students. This is very important for professional work but many students do not know about it until they have their first real jobs.

Commented [10R9]: Ok

Commented [11]: Several other risks:
our research team is trying to publish papers on interpretability/explanability and write grant proposals on equity (fairness/trust/bias). There are long term goals and deadline beyond your work. Although I would love for you to stay forever; how can we still use and build upon the system/code once you are gone? Please state some process/maintenance/documentation to ensure smooth usage/transition and guard against junior undergrads who might break things.

Commented [12R11]: @choupeiyuan.ca@gmail.com please comment since we will be maintaining this system and research after Mike finishes.

Commented [13R11]: Excellent point. I've added in a risk section for project continuity. I will make sure that there is sufficient knowledge transfer that you are not slowed down.

Conclusion

The proposed solution addresses the core requirements of implementing various model interpretability options as well as creates the groundwork to allow for distributed development and integration of further interpretability requirements. As the overall research project has a longer time horizon than this sub-project, this solution will supply both the interpretability implementations as well as a solid foundation for future collaborative interpretability development.