A User-Centered Design of Explainable AI for Clinical Decision Support

Mozhgan Salimiparsa^{†,*}, Daniel Lizotte[†], Kamran Sedig[†]

[†] University of Western Ontario

1. Background

Clinical decision support (CDS) systems are computer applications whose goal is to facilitate the decision-making process of clinicians [1, 2]. In recent years, interest has grown around applying Machine Learning (ML) to make predictions related to clinical outcomes for use in CDS systems [3] by constructing models from patient data. However, predictions made by ML models are often not easily interpretable [4], in the sense that the end-user is often not able to understand why each prediction is made. Particularly in clinical applications, where interpretability and trustworthiness of a model are as important as its accuracy [5], the limited interpretability of many ML models is a major barrier to adoption [6]. This barrier has sparked research interest in interpretable and explainable AI methods, commonly known as XAI. XAI methods are used to construct and communicate explanations of the predictions made by machine learning models [7] so that end-users can interpret those predictions. Such methods have been developed to facilitate interpretability and thereby address barriers to adoption stemming from accountability and trustworthiness [8], and several have been applied to CDS systems. However, these methods are not designed based on end-users' needs; rather, they are based on the developers' intuitions of what a good explanation is [9]. Furthermore, XAI methods are not tailored to the specific tasks that a user will undertake, nor are they tailored to the interface used to perform those tasks. We hypothesize that by explicitly designing XAI methods based on end-user needs and based on the relevant tasks and interfaces, we will be able to provide the interpretability needed to overcome barriers to adoption of ML in CDS systems.

2. Objective

To tackle these issues, we propose to develop a visual analytic tool to explain an ML model for clinical application whose design will explicitly take into account the context of tasks and the needs of end-users. The class of ML models we have chosen to explain is the hybrid feature- and kernel-based predictive model class (HFK) which is derived from Kriging method [10]. The visual analytic tool will explain how the model incorporates both feature-based and similarity-based information to make predictions in order to make the operation of the model more interpretable. We consider three different user categories with different explanations: clinicians, health researchers, and developers. Clinicians and health researchers require explanations about the reasoning behind predictions for a specific patient and a patient population, respectively. Developers require detailed technical information about how predictions are made for both individual and population levels in order to iterate and improve the model. Therefore, by targeting the users and their needs, the main anticipated outcome of this research will be improved usability, utility of the tool in clinical settings, and improved model performance over time.

* msalimip@uwo.ca

3. Literature Review

Various XAI methods have been applied in a variety of clinical settings. Knowledge distillation methods compress a machine model to make it simpler and therefore easier for humans to understand. Such methods have been applied to explain predictions of mortality and ventilator-free days [11], and they have been used to learn interpretable features from deep learning models on a real-world clinical time-series dataset [12]. Several approaches for explaining models in terms of feature importance have been applied to clinical problems including predicting stroke outcome [5], predicting ICU mortality [13], and detecting important clinical features within a large EHR dataset [14]. Layer Relevance Propagation (LRP) is a technique for explaining deep computer vision models wherein a heatmap with an indication of the relevance of pixels in the outcome is presented. LRP has been used to explain clinical events predictions and survival time [15], clinical gate analysis [16], and Alzheimer’s disease [17].

Visual analytic tools have been used to open the black box and explain the logic of the ML models [18–21]. However, the application of visual analytics tools for XAI in CDSS has been much more limited. The RetainVis tool was developed to explain RNNs that make predictions from MRI data [22]. This tool provides a combination of explanations, such as providing displaying outputs change according to input changes, contribution score of each input in the decision, and what-if case analysis. Another visual analytic tool used case-based reasoning to show similar cases for a specific patient with breast cancer [23]. It used a rainbow box to show a comparison of drug properties and a table to describe details about similar cases. A third tool called RuleMatrix has been proposed to explain neural network behavior. This tool provides users with a rule-based explanation and data filtering capabilities, and its use was exemplified with two clinical usage scenarios of cancer and diabetes classifiers [24].

4. Methodology

Our design of a human-centered visual analytic tool will have two main phases. We will involve users in both phases of choosing a XAI method and designing the interface as follows.

1) Choosing an XAI method: As mentioned in the literature review, many XAI methods have been proposed. Each of these methods explains ML models’ predictions in a specific way. To select appropriate XAI methods and explanations based on users’ needs, we adopt Wang et al.’s [25] framework, which is based on human reasoning. This framework suggests how to decide which XAI methods can satisfy end-users’ reasoning goals. In this phase, end-users fill a questionnaire and based on the result, XAI method are chosen.

2) Designing the interface for presenting the explanation: To design the interface, we define and describe all activities that the tool will support. These activities are composed of tasks such as organizing, ordering, summarizing, querying, locating, and clustering. These tasks are further divided into multiple lower-level tasks until they can be supported by visual or interactive tasks such as aggregating, aligning, identifying, and ranking [26]. Tasks and sub-tasks are combined to support users to reach their overall goal. Indeed, all of the visual sub-tasks and low-level interaction that a user performs give emergence to tasks, and sequences of tasks would lead to accomplishing an activity. To assign visual properties to these visual and interactive tasks, we will ask users about visualization tools that they are familiar with and also their preference for visual properties for each visual task.

To this date, I have examined the properties of the ML model and have explored different XAI methods. Also, I have investigated the required information to prepare the questionnaire for phase one and the interview questions for phase two. In the next step, we will contact prospective users to do the study for these two phases.

References

- [1] P. N. Finlay. *Introducing Decision Support Systems - Paul N. Finlay - Google Books*. 1994. URL: https://books.google.ca/books/about/Introducing{_}Decision{_}Support{_}Systems.html?id=RhBPAAAAAAAJ{\&}redir{_}jesc=y (visited on 07/31/2020).
- [2] J. D. Moon and M. Galea. “Overview of clinical decision support systems in healthcare”. In: *Medical Imaging: Concepts, Methodologies, Tools, and Applications* (2016), pp. 1501–1527. DOI: [10.4018/978-1-5225-0571-6.ch064](https://doi.org/10.4018/978-1-5225-0571-6.ch064).
- [3] F. Shamout, T. Zhu, and D. A. Clifton. “Machine learning for clinical outcome prediction”. In: *IEEE Reviews in Biomedical Engineering* (2020).
- [4] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. “Explaining explanations: An overview of interpretability of machine learning”. In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.
- [5] E. Zihni, V. I. Madai, M. Livne, I. Galinovic, A. A. Khalil, J. B. Fiebach, and D. Frey. “Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome”. In: *Plos one* 15.4 (2020), e0231166.
- [6] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. “What clinicians want: contextualizing explainable machine learning for clinical end use”. In: *Machine Learning for Healthcare Conference*. PMLR. 2019, pp. 359–380.
- [7] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58 (2020), pp. 82–115.
- [8] A. Adadi and M. Berrada. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6 (2018), pp. 52138–52160. ISSN: 21693536. DOI: [10.1109/ACCESS.2018.2870052](https://doi.org/10.1109/ACCESS.2018.2870052).
- [9] T. Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial intelligence* 267 (2019), pp. 1–38.
- [10] M. A. Oliver and R. Webster. “Kriging: a method of interpolation for geographical information systems”. In: *International Journal of Geographical Information System* 4.3 (1990), pp. 313–332.
- [11] Z. Che, S. Purushotham, R. Khemani, and Y. Liu. “Interpretable deep models for ICU outcome prediction”. In: *AMIA annual symposium proceedings*. Vol. 2016. American Medical Informatics Association. 2016, p. 371.
- [12] Z. Che, S. Purushotham, R. Khemani, and Y. Liu. “Distilling knowledge from deep networks with applications to healthcare domain”. In: *arXiv preprint arXiv:1512.03542* (2015).
- [13] G. J. Katuwal and R. Chen. “Machine learning model interpretability for precision medicine”. In: *arXiv preprint arXiv:1610.09045* (2016).
- [14] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun. “Retain: An interpretable predictive model for healthcare using reverse time attention mechanism”. In: *arXiv preprint arXiv:1608.05745* (2016).
- [15] Y. Yang, V. Tresp, M. Wunderle, and P. A. Fasching. “Explaining therapy predictions with layer-wise relevance propagation in neural networks”. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE. 2018, pp. 152–162.
- [16] F. Horst, D. Slijepcevic, S. Lapuschkin, A.-M. Raberger, M. Zeppelzauer, W. Samek, C. Breiteneder, W. I. Schöllhorn, and B. Horsak. “On the Understanding and Interpretation of Machine Learning Predictions in Clinical Gait Analysis Using Explainable Artificial Intelligence”. In: *arXiv* (2019).
- [17] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter. “Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer’s disease classification”. In: *Frontiers in aging neuroscience* 11 (2019), p. 194.
- [18] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu. “Understanding hidden memories of recurrent neural networks”. In: *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2017, pp. 13–24.
- [19] Q. Wang, Y. Ming, Z. Jin, Q. Shen, D. Liu, M. J. Smith, K. Veeramachaneni, and H. Qu. “Atmseer: Increasing transparency and controllability in automated machine learning”. In:

- Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–12.
- [20] Q. Li, K. S. Njotoprawiro, H. Haleem, Q. Chen, C. Yi, and X. Ma. “Embeddingvis: A visual analytics approach to comparative network embedding inspection”. In: *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2018, pp. 48–59.
 - [21] X. Zhao, Y. Wu, D. L. Lee, and W. Cui. “iForest: Interpreting random forests via visual analytics”. In: *IEEE transactions on visualization and computer graphics* 25.1 (2018), pp. 407–416.
 - [22] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. “Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records”. In: *IEEE transactions on visualization and computer graphics* 25.1 (2018), pp. 299–309.
 - [23] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi. “Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach”. In: *Artificial intelligence in medicine* 94 (2019), pp. 42–53.
 - [24] Y. Ming, H. Qu, and E. Bertini. “Rulematrix: Visualizing and understanding classifiers with rules”. In: *IEEE transactions on visualization and computer graphics* 25.1 (2018), pp. 342–352.
 - [25] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. “Designing theory-driven user-centric explainable AI”. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–15.
 - [26] S. S. Abdullah. “Visual Analytics of Electronic Health Records with a focus on Acute Kidney Injury”. In: (2020).