AI-based Visual Support for Clinical Diagnosis of Pediatric Suprasellar Tumors and Impacts on Decision-Making Confidence

Eric W. Prince* — University of Colorado

David M. Mirsky† — Children's Hospital Colorado

Todd C. Hankinson‡ — Children's Hospital Colorado

Carsten Görg§ — Colorado School of Public Health

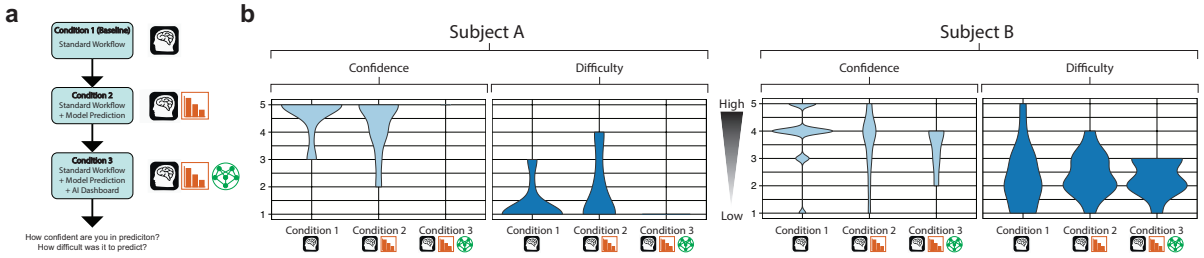

Figure 1: Study overview. a. Experimental design. b. Selected primary results.

Abstract

Explaining AI-based predictions is fundamental for the development of clinical decision support systems. A common visual approach for explaining imaging data predictions is to overlay saliency maps onto images to allow users to interpret what visual features are associated with a given prediction. This approach can be difficult to utilize when differentiating nuanced concepts. For example, clinicians in neuro-oncology will commonly have to differentiate between a group of similar brain tumors (i.e., a radiographic differential diagnosis). We hypothesized that visual representations of counterfactual conditions could improve the utility of AI-based predictions in the context of such a clinical task because it is analogous to a heuristic commonly used by clinicians when making decisions under uncertainty. We present an initial pilot study in which two board-certified clinicians participated in a three condition study to explore how counterfactual visualizations impact diagnostic performance, decision-making confidence, and decision-making difficulty.

1 Introduction

Various visualization approaches already contribute to critical components of clinical decision support systems related to the treatment of patients with brain tumors. The most widely used and essential examples are visual representations of electronic health records and clinical imaging rendering [3]. Artificial intelligence (AI) is expected to make a significant contribution to clinical decision-making related to the treatment of patients with brain tumors by extending the functionality of current visualization systems with predictive inference. However, very few examples exist of this technology within a real-world clinic, in part due to the lack of trust in black-box AI models. In response to this and related challenges in other domains, the VIS community has begun to investigate new methods in the multi-disciplinary field of explainable AI (XAI) [1].

XAI is a rapidly developing field, and visualization plays a key role as the communication bridge between the machine and user. Guidelines have been developed by the United States Federal Drug Administration, the European Union's General Data Protection Regulation, and others which emphasize the use for XAI in decision support AI software. The most commonly applied XAI methods for imaging data are saliency maps generated by methods such as SHAP or LIME [2, 5]. These methods are visually interpretable in scenarios where enough semantic difference exists across images. For example, in images of dogs versus cats there are clear visual features that exist to differentiate the two animals. However, in the context of differentiating suprasellar brain tumors, the visual features become more difficult to confidently interpret, see Figure 2.

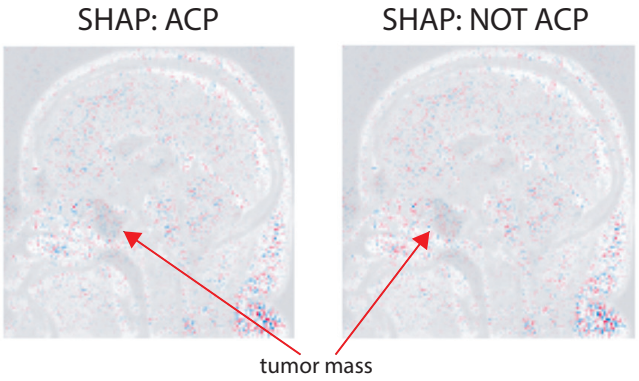

Figure 2: Example of SHAP-generated saliency map for preoperative MR image classifier with 80% test accuracy. This demonstrates that the saliency map is not visually interpretable and does not explain why this classifier correctly classifies the image as ACP.

Inspired by currently used heuristics in clinical decision making, we hypothesized that clinicians will be able to make diagnostic predictions more confidently, with less difficulty, and with greater accuracy if they are able to query *"What are the most similar and dissimilar previously seen patients?"* when given a novel case to diagnose. This concept is known as *representativeness*; a heuristic that clinicians use to interpret diagnostic data by considering similarity of a single case to a group of previously seen cases [6]. This method helps clinicians to make decisions under uncertainty, but is prone to bias and can be inconsistent between clinicians. XAI may provide a solution to answering this question more objectively and support this decision making task.

*e-mail: Eric.Prince@CUAnschutz.edu

†e-mail: David.Mirsky@ChildrensColorado.org

‡e-mail: Todd.Hankinson@ChildrensColorado.org

§e-mail: Carsten.Goerg@CUAnschutz.edu

We have previously demonstrated the feasibility of predicting binary diagnosis of a rare brain tumor versus other brain tumors commonly found in a radiographic differential diagnosis [4]. The purpose of this pilot study was to extend that work by conducting a user study using our previously developed model in combination with a counterfactual matching visualization tool.

2 Methods

We utilized a dataset of preoperative Magnetic Resonance (MR; $n = 52$) and Computed Tomography (CT; $n = 61$) image volumes [4] and followed a human-centered design approach. We met with a board-certified neurosurgeon and neuroradiologist to conduct interviews regarding their workflow for diagnosing suprasellar tumors. We surveyed the literature to derive a library of visualization methods and discussed which techniques would be most relevant. We utilized the ICE-T evaluation framework [8] for our study to assess the visualization's value for our specific domain.

We repurposed Google's What-If Tool (WIT) for counterfactual explanations [9]. Briefly, the WIT identifies counterfactuals using L_1 or L_2 norms in the output layer of a TensorFlow model. Typically, the WIT provides the functionality of modifying inputs; enabling the user to query *"What if I change this feature value?"* We disabled this functionality for our purposes in order to simplify the interface. Instead, users were only able to observe what similar and dissimilar previously seen patients were present in the dataset, thus providing the functionality of counterfactual matching, see Figure 3.

We conducted a three condition study (Figure 1a) with our two expert subjects, each subject performed all three conditions. For all study conditions, subjects were given a set ($n = 28$) of interactive PDF documents which linked to the Open Health Imaging Foundation (OHIF) Viewer [7]. The OHIF viewer provides a standard radiographic interpretation framework. Subjects were tasked with binary diagnostic prediction of adamantinomatous craniopharyngioma (ACP) versus other suprasellar tumors (NOTACP). In addition, subjects would respond to *How confident are you in your prediction?* and *How difficult was your decision?* using a 5-point Likert scale. For each patient, there was also a free response field for subjects to provide any additional feedback. The first condition provided only the OHIF Viewer. The second condition extends the first condition with a predicted value. Predictions were generated using a deep learning model previously published [4]. The third condition extends the second condition with the What-If Tool. Subjects were prompted to think aloud during each condition and audio and screen recordings were captured for each session.

3 Results

There was no significant change in decision-making confidence and difficulty for each subject across the three study conditions for the *NOTACP* class of data (data not shown). However, there was a trend for increased diagnostic confidence and decreased diagnostic difficulty for both subjects with predictions for the ACP class of data (Figure 1b). This trend was strongest for the third condition of the study (OHIF Viewer + AI Prediction + WIT). Finally, there was no significant change in diagnostic accuracy, sensitivity and specificity with respect to each subject across the three study conditions (data not shown).

4 Discussion

Using the WIT, both subjects utilized our intended functionality of the WIT which was to ask: *"What are the most similar and dissimilar previously seen patients?"* In the course of querying this information using the WIT, subjects positively remarked that the counterfactual reasoning was very medically reasonable. Unfortunately, the WIT was very difficult for users to engage with; a significant amount of time was required to explain the concept of the WIT as well as the software interface. This highlights that counterfactual based explanations are likely useful, but there is a critical need for design of new user interfaces which are accessible to clinical users.

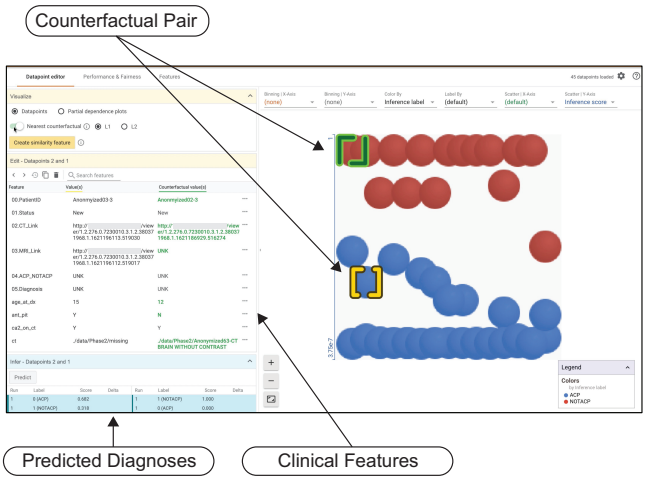

Figure 3: Overview of WIT showing AI predictions, associated clinical features, and matched pairs. Each dot on the scatter plot represents a single patient and is colored by the predicted class.

Our preliminary study has multiple limitations associated with the binary classification task not accurately reflecting a real clinical diagnosis and the use of a small single institutional dataset. Future work will explore these limitations by performing similar studies with a multi-class classifier and a larger subject group who have not previously seen the clinical images.

References

[1] J. Choo and S. Liu. Visual analytics for explainable deep learning. *IEEE computer graphics and applications*, 38(4):84–92, 2018.

[2] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[3] G. A. Norris, J. Garcia, T. C. Hankinson, M. Handler, N. Foreman, D. Mirsky, N. Stence, K. Dorris, and A. L. Green. Diagnostic accuracy of neuroimaging in pediatric optic chiasm/sellar/suprasellar tumors. *Pediatric Blood & Cancer*, 66(6):e27680, June 2019. doi: 10.1002/pbc.27680

[4] E. W. Prince, R. Whelan, D. M. Mirsky, N. Stence, S. Staulcup, P. Klimo, R. C. Anderson, T. N. Niazi, G. Grant, M. Souweidane, et al. Robust deep learning classification of adamantinomatous craniopharyngioma from limited preoperative radiographic images. *Scientific reports*, 10(1):1–13, 2020.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[6] M. Richie and S. A. Josephson. Quantifying heuristic bias: anchoring, availability, and representativeness. *Teaching and learning in Medicine*, 30(1):67–75, 2018.

[7] T. Urban, E. Ziegler, R. Lewis, C. Hafey, C. Sadow, A. D. Van den Abbeele, and G. J. Harris. Lesiontracker: extensible open-source zero-footprint web viewer for cancer imaging research and clinical trials. *Cancer research*, 77(21):e119–e122, 2017.

[8] E. Wall, M. Agnihotri, L. Matzen, K. Divis, M. Haass, A. Endert, and J. Stasko. A heuristic approach to value-driven evaluation of visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):491–500, 2018.

[9] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.