Intrinsically Interpretable Artificial Neural Networks for Learner Modeling

Juan D. Pinto
University of Illinois
Urbana-Champaign
jdpinto2@illinois.edu

Luc Paquette
University of Illinois
Urbana-Champaign
lpaq@illinois.edu

Nigel Bosch
University of Illinois
Urbana-Champaign
pnb@illinois.edu

ABSTRACT

Modern AI algorithms are so complex, that it is often impossible even for expert AI engineers to fully explain how they make decisions. Researchers in education are increasingly using such "black-box" algorithms for a wide variety of tasks. This lack of transparency has rightfully raised concerns over issues of fairness, accountability, and trust. Post-hoc explainability techniques exist that aim to address this issue. However, studies in both educational and non-educational contexts have highlighted fundamental problems with these approaches. In this proposed project, we take an alternative approach that aims to make complex AI learner models more *intrinsically* interpretable, while illustrating how such interpretability can be evaluated. We aim to (1) develop an interpretable neural network, comparing accuracy and issues relevant to interpretability approaches as a whole, (2) evaluate this model's level of interpretability using a human-grounded evaluation approach, and (3) validate the model's inner representations and explore some hypothetical advantages of interpretable models, including their use for knowledge discovery.

Keywords

Explainable AI, evaluating interpretability, model transparency, interpretable neural networks

1. INTRODUCTION

As machine learning models become more powerful and complex, they can become difficult or impossible to interpret, making their decision-making process opaque. When used in high-stakes domains like education, this lack of interpretability raises concerns around fairness, accountability, and trust [7], while also obscuring pedagogical insights that could improve learning outcomes among students. Most efforts to tackle this issue rely on post-hoc techniques for generating explanations of a model's inner workings. However, these methods have shown serious inherent limitations, making them inappropriate for use in educational settings.

J. Pinto, L. Paquette, and N. Bosch. Intrinsically interpretable artificial neural networks for learner modeling. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 982–985, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12730021

The project proposed in these pages addresses this problem by developing and evaluating an *intrinsically interpretable* model—one in which the decision-making process is *interpretable by design*, thereby making problematic post-hoc explanations unnecessary.

This research will be conducted across three main studies. The first will be focused on developing and improving an interpretable neural-network-based learner model, the second will seek to evaluate the intrinsic interpretability of such a model, and the third will validate the patterns identified by this model. This research is guided by the following overarching questions:

1. What tradeoffs exist between accuracy and interpretability when developing an interpretable learner model?
2. How can the interpretability of such a model be evaluated?
3. How can the validity of such a model be evaluated, and what implications does interpretability have for knowledge discovery?

2. BACKGROUND AND RELATED WORK

The growing use of opaque "black-box" models in education has been called the *challenge of interpretability* [1]. Such models are commonly used to trace learners' knowledge (detect what they know) [14], identify affective states (such as boredom, confusion, engagement) [9], or predict student success [5]. However, their lack of transparency can impede pedagogical usefulness, raise ethical concerns around accountability, and erode trust among stakeholders (such as students, teachers, or parents).

One way to address this challenge is though a post-hoc explanation approach. These techniques, borrowed from the broader eXplainable AI (XAI) community, are applied after a model has been created (hence the name) and try to extract insights from indirect observations of model predictions. Most post-hoc methods are model agnostic and rely exclusively on inputs and outputs [3]. Two of the most commonly used post-hoc methods are Local Interpretable Model-agnostic Explanations (LIME) [15] and SHapley Additive exPlanations (SHAP) [11]. In both cases, a surrogate interpretable model is the key behind the explanations.

However, such post-hoc methods have demonstrated problematic limitations, including disagreement between

techniques [8], the risk of generating unjustified counterfactual explanations [10], and the "blind" assumptions required when treating a model as a literal black box [16]. Rudin [16] goes as far as to suggest that post-hoc approaches are too problematic to be used in high-stakes scenarios and that we should instead be focused on trying to create intrinsically interpretable models. Only recently have researchers begun to bring awareness of such problems to education [20, 18]. Echoing Rudin [16], Swamy, Frej, et al. [19] have called for a shift in focus from post-hoc explanations to intrinsically interpretable models within the educational data mining community.

Along this vein, the current proposal serves as a test case for both the development and evaluation of an educational model that is interpretable by design.

3. METHODOLOGY

Our approach to interpretability involves constraining the inherent flexibility of a complex model (specifically a convolutional neural network, or CNN) designed to detect a particular learning behavior (gaming the system) so that it better approximates the way humans understand this behavior. This general line of thinking has been previously explored [9, 17, 21], though the implementation details vary significantly. Building on a prototype model that achieved accuracy on par with a rule-based expert model of gaming-the-system behavior [13], the first study in this project seeks to further develop multiple interpretable CNNs at increasing levels of complexity for later evaluation.

3.1 Developing an Interpretable Neural Network

Gaming the system is a well-studied learner behavior defined as "attempting to succeed in an interactive learning environment by exploiting properties of the system rather than by learning the material" [2]. We chose to focus on detecting this particular behavior due to existing literature, datasets, and models. Specifically, we are using data previously reported in Paquette et al. [12], consisting of sequences of actions from 59 students using the Cognitive Tutor Algebra system during an entire school year. Cognitive Tutor tasks students with solving multi-step mathematical problems and can provide on-demand hints. A total of 10,397 clips (i.e., student action subsequences) were previously labeled by an expert [2] and contained 708 examples of gaming-the-system behavior (6.8%). Using this dataset allows us to directly compare the accuracy of our CNNs with the model based on human expert insights described by Paquette et al. [12].

The models we are developing make it possible to interpret their own convolutional filters, which essentially contain the gaming-the-system patterns identified by the model. This interpretability is achieved by constraining the weights of the network's convolutional filters to be binary rather than continuous—an approach first described by Jiang & Bosch [6]. Since each learned filter serves as a different pattern that the network is seeking in the data, making filter weights binary makes it far easier to understand which combinations of student actions the model considers important. Binary filters also align more closely with the binary nature of the input variables used by the model, further improving interpretability.

This first study will pursue further refinements and model fine-tuning. Modifications to our prototype model include adjusting the feature set, constraining the number of activated features, incorporating variable sequence lengths, enforcing valid feature combinations, and explicitly matching convolutional filters with inputs. This will allow us to tackle some critical questions that remain unanswered, including the tradeoffs between allowing the network too much or too little flexibility, as well as the impact that additional constraints have on model accuracy and interpretability. Addressing these issues further requires confronting the question of how to evaluate interpretability, which is the focus of our second study.

3.2 Evaluating Interpretability

While the field has devised and validated robust methods for measuring different aspects of model *accuracy*, evaluation methods for *interpretability* have yet to be agreed-upon. Within the framework proposed by Doshi-Velez & Kim [4], this project takes what they call a *human-grounded* evaluation approach. This category sits between application- and functionally-grounded evaluation, reaching a sound middle-ground that directly measures a model's interpretability for real humans, but through a simplified task that may not capture the specific needs of learners themselves.

While an application-grounded evaluation approach may provide stronger evidence of interpretability for end-users (ie. students/teachers), its narrow scope would present challenges beyond the increased resources required. Finding that a model is not interpretable for end-users would not inform us of its interpretability for other stakeholders, such as researchers. In this sense, a human-grounded approach can serve as a precursor to more specific application-grounded evaluation. The particular interpretability approach we are taking also has the added benefit of being easy to translate to student-teacher-friendly visualizations, so the distinction between the two evaluation approaches in this case is rather minor. A human-grounded approach, on the other hand, will allow us to evaluate the interpretability of explanations at multiple levels of complexity, all while still maintaining the accuracy of such explanations in regards to the models' inner workings (in contrast to unreliable post-hoc explanations).

For this evaluation, we will assign participants two tasks designed to measure how well they have understood the model's decision-making process. These are (1) forward simulation, in which participants must correctly predict the model's output given specific inputs and (2) counterfactual simulation, in which participants must correctly identify how a specific input needs to be changed in order to alter the model's given output. Participants from a wide range of backgrounds (both with and without prior experience in machine learning) will perform these tasks through a questionnaire. As part of an iterative pilot phase, we will first test the questionnaire with a group of graduate students in an instructional technology program. This will allow us to identify potential issues with its questions or organization. we will then seek participants through various sources, including the mailing list of the International Educational Data

Mining Society (IEDMS) and the newsletter of the Society for Learning Analytics Research (SoLAR).

The questionnaires will be administered through a digital platform. It will include a consent form, questions regarding participants' background (to gather demographic data and level of experience in related disciplines), an explanation of the project and the task, an example of the task, a practice task, and finally, the tasks themselves. For both tasks, participants will be provided with a set of the model's convolutional filters (which contain the gaming-the-system patterns identified by the model).

The forward simulation task will present participants with the inputs for a particular instance—that is, the values for each variable for a given clip of student actions. Participants will be asked to predict whether the model would label this clip as *gaming the system* or *not gaming the system*, given the patterns in the convolutional filters. They will be asked to perform this task five times, receiving a new set of inputs for a different clip each time.

The counterfactual simulation task will again present participants with the inputs for a specific clip, but this time also providing the model's predicted label. Participants will be asked to identify a single change to the inputs that would alter the model's prediction. For example, given a series of inputs and the model's prediction of *not gaming the system*, what change to the inputs would result in the model labeling this clip as *gaming the system*. Participants will select the single correct answer from a series of multiple choice options. As with the first task, this will also be repeated five times with five different scenarios.

After these tasks are complete, participants will be asked to answer a series of questions regarding how they tackled their tasks and how confident they felt about their responses. These confidence levels will allow us to explore how well one's perception of model interpretability aligns with actual interpretability.

For both the forward and counterfactual simulations, participants' accuracy rate (proportion correct out of total questions) will be calculated. The mean accuracy rate for the entire sample will be used as the principal measure of the model's actual interpretability. With a goal of 100 participants, we expect to reach an estimated accuracy rate with a 95% confidence level and a margin of error $< 10\%$. The Cohen's Kappa (inter-rater reliability) will also be calculated to determine the level of agreement between participants. Participants' responses to the multi-class problems (those in the counterfactual simulation) will first be converted to correct/incorrect in order for Kappa to measure general agreement between these two classes. To compare the accuracy rates between participants with prior ML experience and those without, an independent-samples t-test will be used.

To explore the interplay between complexity and interpretability, we will ensure that the number of convolutional filters provided to participants for their tasks vary. If they can accurately perform the tasks with less filters but make more errors when presented with more filters, this will be an indication that increased complexity hinders interpretability.

For a more in-depth analysis, we will create a confusion matrix from responses to the binary problems (those in the forward simulation). Along with calculations of sensitivity (recall), specificity, and precision, this will allow us to dive into specific areas where participants may have encountered trouble. For multi-class problems, a contingency table will serve a similar purpose and help us identify if specific questions or answer choices were particularly difficult for participants.

3.3 Model Validation

Our third study will be an analysis of the specific patterns identified automatically by the model. Here we seek to validate whether the model in fact aligns with reality or if the patterns it picks out are nonsensical within the context of the data and experts' understanding of gaming-the-system behavior. To do this, we will conduct both a quantitative analysis of the patterns themselves and a qualitative interview of an expert in this particular learner behavior.

The quantitative analysis will involve comparing the patterns identified by the model with those identified by the expert as published in Paquette et al. [12]. We will do a pairwise comparison of patterns by calculating the Euclidean distance, cosine similarity, and matrix correlation between any two patterns. These metrics will allow us to identify the most similar patterns between the two models. For the Euclidean distance, we will use the Frobenius norm for this rather than the nuclear norm in order to consider all the elements of the patterns equally.

The purpose of this quantitative analysis is to get a general sense of how similar the CNN's patterns are to those identified by the expert. On one hand, if the model is able to identify similar patterns, this will provide evidence that it is indeed picking up on gaming-the-system behavior. On the other hand, having some more unique patterns as well may indicate that the model is able to identify novel behavior patterns that are not obvious to humans. However, to test whether this latter case is true and it is not simply the model picking up on non-sensical patterns, we will conduct an interview with an expert in gaming-the-system behavior.

In this interview, we will ask this expert to engage in a think-aloud as they are presented with each pattern identified by our model, and they will explain whether the pattern makes sense for gaming-the-system detection or why not. We will also ask them to provide their overall impression of the model, its accuracy, and its interpretability. They will be asked whether and how they might see such a model being used to inform theories of learning or to inform our understanding of the learner behavior being modeled. Together with the results of the quantitative analysis, the interview will provide an overall sense of the model's validity to real-world student action patterns indicative of gaming the system.

4. CONCLUSION

While this project focuses on one specific type of algorithm (convolutional neural networks) designed for a particular

task (detection of student gaming-the-system behavior) using a single interpretability approach (constraining the complexity of the convolution filters), its significance lies in deeply expanding the conversation in a direction that is still relatively unexplored within the educational data mining community. The expected outcomes of this project include:

1. An intrinsically interpretable neural network model for detecting a learner behavior, aligned with expert understandings and optimized for performance and interpretability.
2. A robust approach for evaluating the interpretability of neural network models in educational contexts.
3. Insights into the implications of intrinsic interpretability for evaluating model validity and knowledge discovery in education.
4. Contributions to addressing the challenge of interpretability in educational data science, with potential implications for fairness, accountability, trustworthiness, and pedagogical usefulness.

Overall, this research aims to advance the development and evaluation of interpretable machine learning models in education, promoting transparency and accountability in AI-enabled educational technologies.

5. REFERENCES

[1] R. S. Baker. Challenges for the future of educational data mining: The baker learning analytics prizes. *Journal of Educational Data Mining*, 11(1), 2019.

[2] R. S. Baker and A. M. J. A. de Carvalho. Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining (EDM)*, pages 38–47, 2008.

[3] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), July 2019.

[4] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, Mar. 2017.

[5] Q. Hu and H. Rangwala. Academic performance estimation with attention-based graph convolutional networks. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, pages 69–78, 2019.

[6] L. Jiang and N. Bosch. Predictive sequential pattern mining via interpretable convolutional neural networks. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, pages 761–766, 2021.

[7] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, Jan. 2022.

[8] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective, Feb. 2022.

[9] A. S. Lan, A. Botelho, S. Karumbaiah, R. S. Baker, and N. Heffernan. Accurate and interpretable sensor-free affect detectors via monotonic neural networks. In *Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20)*, 2020.

[10] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations, July 2019.

[11] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[12] L. Paquette, A. M. de Carvalho, and R. S. Baker. Towards understanding expert coding of student disengagement in online learning. In *Proceedings of the 36th Annual Cognitive Science Conference*, pages 1126–1131, 2014.

[13] J. D. Pinto, L. Paquette, and N. Bosch. Interpretable neural networks vs. expert-defined models for learner behavior detection. In *Companion Proceedings of the 13th International Conference on Learning Analytics & Knowledge Conference (LAK23)*, pages 105–107, 2023.

[14] S. Pu and L. Becker. Self-attention in knowledge tracing: Why it works. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, editors, *Artificial Intelligence in Education*, volume 13355, pages 731–736. Springer International Publishing, Cham, 2022.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier, Feb. 2016.

[16] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[17] Y. Shi. Interpretable code-informed learning analytics for CS education. In *Companion Proceedings 13th International Conference on Learning Analytics & Knowledge (LAK23)*, pages 180–185, 2023.

[18] V. Swamy, S. Du, M. Marras, and T. Kaser. Trusting the explainers: Teacher validation of explainable artificial intelligence for course design. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 345–356, Arlington TX USA, Mar. 2023. ACM.

[19] V. Swamy, J. Frej, and T. Käser. The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations, July 2023.

[20] V. Swamy, B. Radmehr, N. Krco, M. Marras, and T. Käser. Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, 2022.

[21] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Chen, Y. Yin, Z. Huang, and S. Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6153–6161, Apr. 2020.