Interpretable Machine Learning in Healthcare

Muhammad Aurangzeb Ahmad
KenSci Inc. Seattle, Washington
Center for Data Science, University of Washington - Tacoma
Seattle, Washington
muhammad@kensci.com

Dr. Carly Eckert M.D*
KenSci Inc. Seattle, Washington
Department of Epidemiology, University of Washington
Seattle, Washington
carly@kensci.com

Ankur Teredesai†
KenSci Inc. Seattle, Washington
Center for Data Science, University of Washington - Tacoma
Seattle, Washington
ankur@kensci.com

ABSTRACT

This tutorial extensively covers the definitions, nuances, challenges, and requirements for the design of interpretable and explainable machine learning models and systems in healthcare. We discuss many uses in which interpretable machine learning models are needed in healthcare and how they should be deployed. Additionally, we explore the landscape of recent advances to address the challenges model interpretability in healthcare and also describe how one would go about choosing the right interpretable machine learnig algorithm for a given problem in healthcare.

CCS CONCEPTS

• **Computing methodologies → Supervised learning**; **Machine learning approaches**; • **Applied computing → Health informatics**;

KEYWORDS

Interpretable Machine Learning, Explainable AI, Machine Learning in Healthcare

ACM Reference Format:
Muhammad Aurangzeb Ahmad, Dr. Carly Eckert M.D, and Ankur Teredesai. 2018. Interpretable Machine Learning in Healthcare. In *ACM-BCB '18: 9th ACM Int'l Conf. on Bioinformatics, Computational Biology, and Health Informatics, Aug. 29–Sept. 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3233547.3233667

1 EXTENDED ABSTRACT

Interpretable Machine Learning refers to machine learning models that can provide explanations regarding why certain predictionsare made. In many domains where user trust in the predictions of machine learning systems is needed, merely providing traditional machine learning metrics like AUC, precision, and recall may not be sufficient. While machine learning techniques have been employed for decades, the expansion of these techniques into fields like healthcare have led to an increased emphasis for explanations of machine learning systems. Clinical providers and other decision makers in healthcare note interpretability of model predictions as a priority for implementation and utilization. As machine learning applications are increasingly being integrated into various parts of the continuum of patient care, the need for prediction explanation is imperative. Machine learning solutions are being used to assist providers across clinical care domains as well as clinical operations, and costs. Decisions based on machine learning predictions could inform diagnoses, clinical care pathways, and patient risk stratification, among many others. It follows, that for decisions of such import, clinicians and others desire to know the "reason" behind the prediction. In this tutorial, we will give an extensive overview of the various nuances of what constitutes explanation in machine learning, explore multiple definitions of explanation.

The contexts within healthcare systems where it may be prudent to ask machine learning systems for explanations vs. explanation agnostic contexts will also be explorred. Thus a physician may be greatly interested in knowing why a machine learning system is suggesting a cancer diagnosis vs. a hospital ED planner would rarely be interested in knowing why a machine learning system is making predictions about hourly arrivals in ED. We also discuss how these definitions map to various machine learning systems and algorithms that are available today - all within a healthcare context. We use results from our research on performance comparison of interpretable models on real world problems like risk of readmission prediction, ED utilization prediction and hospital length of stay prediction to explore the constraints and drivers around going about using explainable machine learning algorithms in various healthcare contexts.

Different aspects of explainability in machine learning will be explored in this tutorial e.g., explainability is not limited to machine learning models but also to other aspects of machine learning like input data, model parameters, and the algorithms used. Additionally, the type of explanation provided to the is highly dependent upon the user of the system e.g., competence (cognitive capacity), novice vs. expert (domain knowledge), depth of explanation (explanation granularity) etc. Thus, in some cases, a simple linear model using highly engineered and complex features may be less interpretable than a deep learning model using simple intuitive features. Based on a comprehensive survey of literature on interpretable machine learning models we describe a framework which can be used to

[*] The secretary disavows any knowledge of this author's actions.
[†] The secretary disavows any knowledge of this author's actions.

ACM-BCB'18, August 28-September 1, 2018, Washington, DC, USA.

ACM ISBN 978-1-4503-5794-4/18/08.
https://doi.org/10.1145/3233547.3233667

evaluate interpretable machine learning systems. We then map this framework and various machine learning algorithms to multiple problem domains within healthcare. We also describe in detail the constraints and pitfalls for interpretable machine learning within healthcare from the perspective of a healthcare domain expert.

In the later half of the tutorial, we focus on real world use cases and studies on machine learning systems in healthcare e.g., A landmark study on a machine learning model predicting pneumonia revealed that the machine learning system was giving a lower risk score to patients who also have asthema. In reality the patients in the asthma cohort were already being given extra care so that the data was biased. Nuances like these get lost in Black Box machine learning systems. In medical image diagnosis where deep learning algorithms have shown to have excellent predictive power, it has been demonstrated that it is possible to fool the system into making mistakes which a human expert would never make. We explore use cases accross the patient care continuum to address the various nuances needed to balance between algorithmic optimization and explanability in problems like disease progression, risk of readmission, emergency department admission and utilization, disease diagnosis etc. Lastly we give our perspective on the future of machine learning and healthcare by extrapolating on some of the current trends, exploring some emerging areas in this field and describing areas where the healthcare domain can benefit the most from application of machine learning.

REFERENCES

[1] Ahmad, Muhammad Aurangzeb, Zoheb Borbora, Jaideep Srivastava, and Noshir Contractor. "Link prediction across multiple social networks." In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, pp. 911-918. IEEE, 2010.

[2] Al-Shedivat, Maruan, Avinava Dubey, and Eric P. Xing. "Contextual Explanation Networks." arXiv preprint arXiv:1705.10301 (2017).

[3] Al-Shedivat, Maruan, Avinava Dubey, and Eric P. Xing. "The Intriguing Properties of Model Explanations."

[4] O. Biran and K. McKeown. Justification narratives for individual classifications. In Proceedings of the AutoML workshop at ICML, volume 2014, 2014.

[5] David Gunning Explainable Artificial Intelligence (XAI) DARPA/I2O 2016

[6] T. Hastie and R. Tibshirani. Generalized additive models . Chapman and Hall/CRC, 1990.

[7] Henelius, Andreas, Kai PuolamÃďki, and Antti Ukkonen. "Interpreting Classifiers through Attribute Interactions in Datasets." (2017).

[8] Denis J Hilton. Conversational processes and causal explanation. Psychological Bulletin, 107(1):65âĂŞ81, 1990.

[9] Kulesza, Todd, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. "Principles of explanatory debugging to personalize interactive machine learning." In Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 126-137. ACM, 2015.

[10] Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490 (2016).

[11] Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate intelligible models with pairwise interactions." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 623-631. ACM, 2013.

[12] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv preprint arXiv:1706.07269 (2017).

[13] Miller, Tim, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of Inmates Running the Asylum." In IJCAI-17 Workshop on Explainable AI (XAI), p. 36. 2017.

[14] Google’s research chief questions value of ’Explainable AI’ George Nott 23 June, 2017 https://www.computerworld.com.au/article/621059/google-research-chief-questions-value-explainable-ai/

[15] Ribeiro, Marco Tulio, Sameer Singh, and Carlos 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. ACM, 2016.

[16] Si, Zhangzhang, and Song-Chun Zhu. "Learning and-or templates for object recognition and detection." IEEE transactions on pattern analysis and machine intelligence 35, no. 9 (2013): 2189-2205.

[17] Ustun, Berk, and Cynthia Rudin. "Supersparse linear integer models for optimized medical scoring systems." Machine Learning 102, no. 3 (2016): 349-391.

[18] Wang, Fulton, and Cynthia Rudin. "Falling rule lists." In Artificial Intelligence and Statistics, pp. 1013-1022. 2015.

[19] M. R. Wick and W. B. Thompson. Reconstructive expert system explanation. Artificial Intelligence, 54(1- 2):33âĂŞ70, 1992.

[20] Yang, Hongyu, Cynthia Rudin, and Margo Seltzer. "Scalable Bayesian rule lists." arXiv preprint arXiv:1602.08610 (2016).