Tutorial on Human-Centered Explainability for Healthcare

Prithwish Chakraborty*†
prithwish.chakraborty@ibm.com
IBM Research
Yorktown Heights, NY, USA

Bum Chul Kwon*
bumchul.kwon@us.ibm.com
IBM Research
Cambridge, MA, USA

Sanjoy Dey*
deysa@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

Amit Dhurandhar*
adhuran@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

Daniel Gruen*
daniel_gruen@us.ibm.com
IBM Research
Cambridge, MA, USA

Kenney Ng*
kenney.ng@us.ibm.com
IBM Research
Cambridge, MA, USA

Daby Sow*
sowdaby@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

Kush R. Varshney*
krvarsh@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

ABSTRACT

In recent years, the rapid advances in Artificial Intelligence (AI) techniques along with an ever-increasing availability of healthcare data have made many novel analyses possible. Significant successes have been observed in a wide range of tasks such as next diagnosis prediction, AKI prediction, adverse event predictions including mortality and unexpected hospital re-admissions. However, there has been limited adoption and use in the clinical practice of these methods due to their black-box nature. A significant amount of research is currently focused on making such methods more interpretable or to make post-hoc explanations more accessible. However, most of this work is done at a very low level and as a result, may not have a direct impact at the point-of-care. This tutorial will provide an overview of the landscape of different approaches that have been developed for explainability in healthcare. Specifically, we will present the problem of explainability as it pertains to various personas involved in healthcare viz. data scientists, clinical researchers, and clinicians. We will chart out the requirements for such personas and present an overview of the different approaches that can address such needs. We will also walk-through several use-cases for such approaches. In this process, we will provide a brief introduction to explainability, charting its different dimensions as well as covering some relevant interpretability methods spanning such dimensions. We will touch upon some practical guides for explainability and provide a brief survey of open source tools such as the IBM AI Explainability 360 Open Source Toolkit.

*authors contributed equally
†corresponding author

KDD '20, August 23–27,2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08.
https://doi.org/10.1145/3394486.3406470

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

explainablility; interpretability; deep learning

ACM Reference Format:
Prithwish Chakraborty, Bum Chul Kwon, Sanjoy Dey, Amit Dhurandhar, Daniel Gruen, Kenney Ng, Daby Sow, and Kush R. Varshney. 2020. Tutorial on Human-Centered Explainability for Healthcare. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020,Virtual Event, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3394486.3406470

TUTORIAL WEBPAGE

The webpage for this tutorial can be accessed at https://healthxaitutorial.github.io/kdd2020/

TARGET AUDIENCE AND PRE-REQUISITE KNOWLEDGE

The primary target for this tutorial are "Data Scientists" interested in building models and systems on healthcare data with improved clinical adoption. The audience is expected to have basic knowledge of machine learning.

TUTORIAL OUTLINE

- Introduction (30 mins)
  - Overview of AI in healthcare
  - Challenges for adoption of AI in practice
- A Brief Intro to Explainability (40 mins)
  - General dimensions of Explainability
  - Open source Tools
  - Model cards and factsheets
  - Some Example Methods
- Need for persona driven Explainability in Healthcare (40 mins)

  - Critique of Explainability
  - General Personas for Explainability
- Persona Specific Perspective (40 mins)
  - Data Scientist perspectives and use-cases
  - Clinical Research perspectives and use-cases
  - Clinician Perspectives and use-cases
- General takeaways for Healthcare personas need for Explainability (20 mins)
- Conclusions and Discussions (30 mins)

SHORT TUTOR BIOGRAPHIES

Prithwish Chakraborty, PhD, is currently a Research Staff Member at IBM Research in the Center of Computational Health at the IBM T.J. Watson Research Center, NY. His work focuses on applications of data science towards patient health characterization and risk modeling.

Bum Chul Kwon is Research Staff Member at IBM Research in the Center of Computational Health. His research area includes visual analytics, data visualization, human-computer interaction, healthcare, and machine learning.

Sanjoy Dey is a Research Staff Member in the Center of Computational Health at the IBM T.J. Watson Research Center, NY. His research area focuses on building diverse machine learning models and applying them in the context of different healthcare and life science problems.

Amit Dhurandhar is a Research Staff Member in the Trusted AI department at IBM T.J. Watson Research NY. He obtained his PH.D. from the University of Florida in 2009. He has worked on projects spanning multiple industries such as Semi-conductor manufacturing, Oil and Gas, Procurement, Retail, Utilities, Airline, Health Care with his current research focus being explainable AI. His recent work was featured in Forbes, PC magazine, New Yorker, Quatrz with corresponding technical contributions in leading research venues such as Science, Nature, NeurIPS, ICML and JMLR.

Dan Gruen is a Cognitive Scientist in IBM Research's Integrated Care Research group in the Center for Computational Health. Dan focuses on AI explainability to support clinical reasoning and patient behavioral change. He also runs IBM's Cognitive Experience Invention Development Team.

Kenney Ng is a Principal Research Staff Member in the Center for Computational Health and manager of the Health Analytics Research Group at IBM Research Cambridge. He received B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology. He is a member of the IEEE and AMIA. His current research focus is on the development and application of data mining, machine learning, and AI techniques to analyze, model and derive actionable insights from real world health data.

Daby Sow is a Principal Research Staff Member at IBM Research in the Center of Computational Health at the IBM T.J. Watson Research Center, NY. He is also a research manager of the Biomedical Analytics and Modeling group within the Center for Computational Health at IBM Research. In this role, he is leading a team of AI scientists developing novel AI solutions for various open healthcare research problems.

Kush R. Varshney is a principal research staff member and manager with IBM Research at the Thomas J. Watson Research Center, Yorktown Heights, NY, where he leads the machine learning group in the Foundations of Trusted AI department. He received the S.M. degree in 2006 and the Ph.D. degree in 2010, both in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge. He is a senior member of the IEEE and a member of the Partnership on AI's Safety-Critical AI expert group.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. https://arxiv.org/abs/1909.03012

[3] Hamsa Bastani, Osbert Bastani, and Carolyn Kim. 2018. Interpreting predictive models for human-in-the-loop analytics. *arXiv preprint arXiv:1705.08504* (2018), 1–45.

[4] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*. 3504–3512.

[5] Sarthak Jain and Byron C Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3543–3556.

[6] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2019. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics* 25, 1 (2019), 299–309.

[7] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[8] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. 2018. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical care medicine* 46, 4 (2018), 547–553.

[9] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1, 1 (2018), 18.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. ACM, 1135–1144.

[11] Cynthia Rudin. 2019. Do Simpler Models Exist and How Can We Find Them?. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1–2.

[12] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206.

[13] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[14] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 572, 7767 (2019), 116.

[15] Xiang Wang, David Sontag, and Fei Wang. 2014. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 85–94.

[16] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 11–20.