Interpretable Machine Learning in Healthcare

Muhammad Aurangzeb Ahmad, Carly Eckert, Ankur Teredesai, and Greg McKelvey

***Abstract*—The drive towards greater penetration of machine learning in healthcare is being accompanied by increased calls for machine learning and AI based systems to be regulated and held accountable in healthcare. Interpretable machine learning models can be instrumental in holding machine learning systems accountable. Healthcare offers unique challenges for machine learning where the demands for explainability, model fidelity and performance in general are much higher as compared to most other domains. In this paper we review the notion of interpretability within the context of healthcare, the various nuances associated with it, challenges related to interpretability which are unique to healthcare and the future of interpretability in healthcare.**

***Index Terms*—Interpretable Machine Learning, Machine Learning in Healthcare, Health Informatics**

I. Introduction

WHILE the use of machine learning and artificial intelligence in medicine has its roots in the earliest days of the field [1], it is only in recent years that there has been a push towards the recognition of the need to have healthcare solutions powered by machine learning. This has led researchers to suggest that it is only a matter of time before machine learning will be ubiquitous in healthcare [22]. Despite the recognition of the value of machine learning (ML) in healthcare, impediments to further adoption remain. One pivotal impediment relates to the *black box* nature, or opacity, of many machine learning algorithms. Especially in critical use cases that include clinical decision making, there is some hesitation in the deployment of such models because the cost of model misclassification is potentially high [21]. Healthcare abounds with possible ”high stakes” applications of ML algorithms: predicting patient risk of sepsis (a potentially life threatening response to infection), predicting a patient's likelihood of readmission to the hospital, and predicting the need for end of life care, just to name a few. Interpretable ML thus allows the end user to interrogate, understand, debug and even improve the machine learning system. There is much opportunity and demand for interpretable ML models in such situations. Interpretable ML models allow end users to evaluate the model, ideally before an action is taken by the end user, such as the clinician. By explaining the reasoning behind predictions, interpretable machine learning systems give users reasons to accept or reject predictions and recommendations.

Audits of machine learning systems in domains like healthcare and the criminal justice system reveal that the decisions and recommendations of machine learning systems may be biased [4]. Thus, interpretability is needed to ensure that such systems are free from bias and fair in scoring different ethnic and social groups [12]. Lastly, machine learning systems are already making decisions and recommendations for tens of millions of people around the world (i.e. Netflix, Alibaba, Amazon). These predictive algorithms are having disruptive effects on society [32] and resulting in unforeseen consequences [12] like deskilling of physicians. While the application of machine learning methods to healthcare problems is inevitable given that complexity of analyzing massive amounts of data, the need to standardize the expectation for interpretable ML in this domain is critical.

Historically, there has been a trade-off between interpretable machine learning models and performance (precision, recall, F-Score, AUC, etc.) of the prediction models [8]. That is, more interpretable models like regression models and decision trees often perform less well on many prediction tasks compared to less interpretable models like gradient boosting, deep learning models, and others. Researchers and scientists have had to balance the desire for the most highly performing model to that which is adequately interpretable. In the last few years, researchers have proposed new models which exhibit high performance as well as interpretability e.g., GA2M [5], rule-based models like SLIM[30], falling rule lists[31], and model distillation [27]. However, the utility of these models in healthcare has not been convincingly demonstrated due to the rarity of their application.

The lack of interpretability in ML models can potentially have adverse or even life threatening consequences. Consider a scenario where the insights from a black box models are used for operationalizating without the recognition that the predictive model is not prescriptive in nature. As an example, consider Caruana *et al.* [5] work on building classifiers for labeling pneumonia patients as high or low risk for in-hospital mortality. A neural network, essentially a black box in terms of interpretability, proved to be the best classifier for this problem. Investigation of this problem with regression models revealed that one of the top predictors was *patient history of asthma*, a chronic pulmonary disease. The model was predicting that given asthma, a patient had a lower risk of in-hospital death when admitted for pneumonia. In fact, the opposite is true - patients with asthma are at higher risk for serious complications and sequelae, including death, from an infectious pulmonary disease like pneumonia. The asthma patients were, in fact, provided more timely care of a higher acuity than their counterparts without asthma, thereby incurring a survival advantage. Similarly leakage from data can misinform models or artificially inflate performance during testing [14], however explanations can be used to interrogate and rectify models when such problems surface.

While there is a call to apply interpretable ML models to a large number of domains, healthcare is particularly challenging due to medicolegal and ethical requirements, laws, and regulations, as well as the very real caution that must be

The authors are from KenSci Inc. Corresponding Author e-mail: (muhammad@kensci.com).

employed when venturing into this domain. There are ethical, legal and regulatory challenges that are unique to healthcare given that healthcare decisions can have an immediate effect on the wellbeing or even the life of a person. Regulations like the European Union's General Data Protection Regulation (GDPR) require organizations which use patient data for predictions and recommendations to provide *on demand* explanations [28]. The inability to provide such explanations on demand may result in large penalties for the organizations involved. Thus, there are monetary as well as regulatory and safety incentives associated with interpretable ML models.

Interpretability of ML models is applicable across all types of ML: supervised learning [17], unsupervised learning [6] and reinforcement learning [15]. In this paper, we limit the scope of the discussion to interpretability in supervised learning models as this covers the majority of the ML systems deployed in healthcare settings [18]. The remainder of the paper is organized as follows: First, we define interpretability in machine learning, we provide an overview of the need for interpretability in machine learning models in healthcare, and we discuss use cases where interpretability is less critical. We conclude this paper with a brief survey of interpretable ML models and challenges related to interpretability unique to healthcare.

II. What is Interpretability?

While there is general consensus regarding the need for interpretability in machine learning models, there is much less agreement about what constitutes interpretability [17]. To this end, researchers have tried to elucidate the numerous notions and definitions of interpretability [17],[8]. Interpretability has been defined in terms of model transparency [17], model fidelity [17], model trust [17], [8], [9], and model comprehension [9], among other characteristics. Many of the notions of interpretability have been developed in the context of computing systems and mostly ignore the literature on interpretability that comes from the social sciences or psychology [19]. Thus, one common objection to these definitions of interpretability is that it does not put enough emphasis on the user of interpretable machine learning systems [16]. This results in a situation where the models and explanations produced do not facilitate the needs of the end users [19].

A primary sentiment of interpretability is the fidelity of the model and its explanation i.e., the machine learning model should give an explanation of why it is making a prediction or giving a recommendation. This is often referred to as a key component of "user trust" [25]. In some machine learning models like decision trees [24], regression models [33], and context explanation networks [3] the explanation itself is part of the model. In contrast, for models such as neural networks, support vector machines, and random forests that do not have explanations as part of their predictions it is possible to extract explanations from models that are applied post-hoc, such as locally interpretable model explanations (LIME) [25] and Shapley Values [26]. LIME constructs explanations by creating a local model, like a regression model, for the instance for which an explanation is required. The data for the local model is generated by perturbing the instance of interest, observing the change in labels and using it to train a new model. Shapley values, on the other hand, take a game theoretical perspective to determine the relative contribution of variables to the predictions by considering all possible combinations of variables as cooperating and competing coalitions to maximize payoff, defined in terms of the prediction [26].

Many definitions of interpretability include transparency of the components and algorithms, the use of comprehensible features in model building, and intelligible applications of parameters and hyperparameters. Based on the work of Lipton *et al.* [17], interpretability can be described in terms of transparency of the machine learning system *i.e.,* the algorithm, features, parameters and the resultant model should be comprehensible by the end user. At the feature level, the semantics of the features should be understandable. Thus, a patient's age is readily interpretable as compared to a highly engineered feature (the third derivative of a function that incorporates age, social status and gender, for example). At the model level, a deep learning model is less interpretable compared to a logistic regression model. An exception to this rule is when the deep learning model utilizes intuitive features as inputs and the regression model utilizes highly engineered features, then the deep learning model may in fact be more interpretable. Lastly, we consider interpretability in terms of the model parameters and hyperparameters. From this perspective, the number of nodes and the depth of the neural network is not interpretable but the number of support vectors for a linear kernel is much more interpretable [17].

Interpretability may also mean different things for different people and in different use cases. Consider regression models. For a statistician or a machine learning expert the following equation for linear regression is quite interpretable:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, ..., n \tag{1}$$

Those familiar with the field, can easily identify the relative weights of the parameter coefficients and abstract meaning from the derived equation. However, most non statisticians, including some clinicians, may not be able to interpret the meaning of this equation. For others, merely describing a model as ”linear” may be sufficient. Conversely, a more advanced audience, knowing the error surface of the model may be needed to consider the model fully "interpretable".

For some predictive algorithms, however, the lack of interpretability may go deeper. Thus consider the following equation for updating weights in a deep learning network.

$$a^l_j = \sigma\Big(\sum_k w^l_{jk} a^{l-1}_k + b^l_j\Big) \tag{2}$$

While the math is clear and interpretable, the equation does not help anyone understand how deep learning networks actually learn and generalize.

Finally, interpretability of machine learning models in healthcare is always context dependent, even to the level of the user role. The same machine learning model may require generating different explanations for different end users e.g., an explanation model for a risk of readmission prediction model to be consumed by a hospital discharge planner vs.

a physician may necessitate different explanations for the same risk score. This component of interpretability parallels the thought processes and available interventions of different personas in healthcare. For example, a discharge planner will often evaluate a patient's risk of readmission based on the components of that patient's situation that are under her purview - perhaps related to the patient's living situation, unreliable transportation, or need for a primary care physician. While the treating physician will need to be aware of these associated characteristics, she may be more likely to focus on the patient's cardiac risk and history of low compliance with medications that are associated with the patient's high risk of readmission. Context is critical when considering interpretability.

III. Interpretability Vs. Risk

While there are a number of reasons why interpretability of ML models is important, not all prediction problems in supervised machine learning predictions require explanations. Alternatives to explanations include domains where the system may have theoretical guarantees to always work or empirical guarantees of performance when the system has historically shown to have great performance e.g., deep learning applications radiology with superhuman performance[20]; or in work pioneered by Gulshan et al, the developed deep learning algorithm was able to detect diabetic retinopathy from retinal fundal photographs with extremely high sensitivity and specificity [10]. The exceptional performance supports the fact that this prediction does not require an explanation. However, findings such as this are quite rare. Another example where interpretability may not be prioritized is in the setting of emergency department (ED) crowding. For a hospital's ED, the number of patients expected to arrive at the ED in the next several hours can be a helpful prediction to anticipate ED staffing. In general, the nursing supervisor is not concerned with the reasons why they are seeing the expected number of patients (of course, there are exceptions) but only interested in the number of expected patients and the accuracy of the prediction. On the other hand, consider the case of predicting risk of mortality for patients. In this scenario, the imperative for supporting explanations for predictions may be great - as the risk score may drive critical care decisions. What these examples demonstrate is that the clinical context (also, how ”close” the algorithm is to the patient) associated with the application determines the need for explanation. The fidelity of the interpretable models also plays a role in determining the need for explanations. Models like LIME [25] produce explanations which may not correspond to how the predictive model actually works. LIME models are post-hoc explanations of model output, and in some ways, likely mimics the manner in which human beings explain their own decision making processes [17], this may be an admissible explanation where explanations are needed but the cost for the occasional false positives is not very high.

Consider Figure 1 which shows a continuum of potential risk predictions related to patient care. The arrow represents the increasing need for explanations along the continuum. Consider a model for cost prediction for a patient, the accuracy of the prediction may take precedence over explanation depending on the user role. However, as we move up the continuum to *Length of hospital stay* explanations may be helpful in decision making while tolerating a slight decrement in model performance. Thus, the specific use case is very important when considering which predictive and explanation models to choose. Certain use cases and domains require us to sacrifice performance for interpretability while in other cases, predictive performance may be the priority.

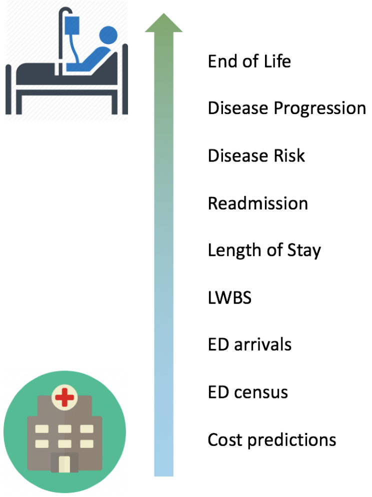

Fig. 1: Prediction Use Cases vs. Need for Interpretability (LWBS: left without being seen)

IV. The Challenge of Interpretability in Healthcare

The motivation for model explanations in healthcare is clear - in many cases both the end users and the critical nature of the prediction demand a certain transparency - both for user engagement and for patient safety. However, merely providing an explanation for an algorithm's prediction is insufficient. The manner in which interpretations are shared with the end users, incorporated into user workflows, and utilized must be carefully considered.

Healthcare workers are generally overwhelmed - by the number of patients they are required to see in a shift, by the amount of data generated by such patients, and the associated tasks required of them (data entry, electronic health record system requirements, as well as providing clinical care). Machine learning algorithms and their associated explanations, if not delivered correctly, will merely be one additional piece of data delivered to a harried healthcare professional. In order to be truly considered, ML output should be comprehensible to the intended user from a domain perspective and be applicable with respect to the intended use case.

A. User Centric Explanations

The participation of end users in the design of clinical machine learning tools is imperative - to better understand how the end users will utilize the output components - and

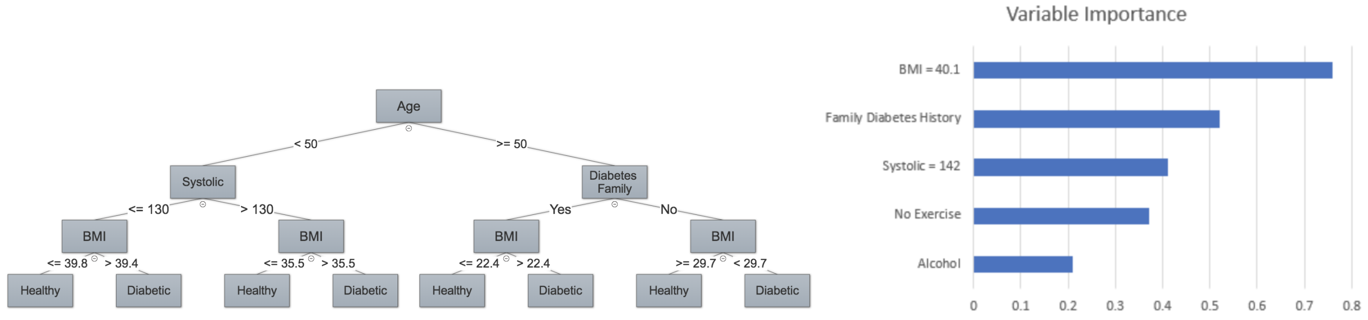

Fig. 2: Global vs. Local Models for Predicting Diabetes

also to educate end users to the nature of the prediction and explanation. According to Jeffrey *et al.* [13], even seasoned clinicians have difficulty interpreting risk scores and probability based estimates and end user input in the design of the expected output can drive participation. Moreso, understanding how end users interpret explanations, derived from the machine learning models, is imperative. Consider for example, the following output:

Patient risk of Readmission: 62, HIGH
Top Factors: Low albumin
Elevated heart rate in emergency department
History of heart failure

How will a healthcare provider interpret this resulting risk score and the associated explanation? Does the fact that the provider knows that these attributes are ”true” for this patient allow user trust in the model? Does the physician consider that by addressing these top factors - such as the patient’s low albumin- that the patient’s risk of readmission will be mitigated? It is important that the concepts of causality and association are emphasized and differentiated. Lipton [17] addresses the issue of algorithm explanation and the tendency to attribute causality to an explanation. He cautions against the conflation of these concepts but does remind us that the results of the explanation can instead inform future formal studies to investigate causal associations between the associated factor and the end point, i.e. readmission.

B. Performance vs. Transparency Trade-off

Earlier in this paper we described the trade-off between model performance and model transparency in healthcare algorithms. How is this trade-off determined? and by whom? Others have described the need to optimize models towards different performance metrics, and that AUC may not always be the metric to optimize. For example, when predicting end of life to determine when to refer patients to hospice, physicians may prefer to optimize for model precision, that is, to maximize the number of individuals who are correctly classified as likely to die by the algorithm. Similarly, the trade-off between performance and interpretability requires discussions with end users to understand the clinical and human risk associated with misclassification or with model opacity.

C. Balancing Requirements of Interpretability

As there is not a single unified conception of interpretability, there are multiple requirements for an ideal interpretable machine learning system, some of which may be at odds. Consider model soundness which refers to how close the model explanation is to how the model actually works. It may be the case that the model which results in the best performance and interpretability is a decision trees with depth 8 and 50 nodes. While the decision tree model is interpretable, the whole model is not comprehensible at the same time. Simultability is a characteristic of a model when it can be comprehensible in its entirety [17]. In this situation, it may be possible to make the decision tree more interpretable by pruning and then use that model for explanations. This may result in a loss in performance and also a loss in soundness, as the model now corresponds to a lesser degree regarding how predictions are being made.

Certain healthcare applications such as predicting disease progression may require explanations at the local, the cohort and the global level. For such applications, local explanations like LIME or Shapley Values may not suffice. One way to address the requirements of explanation scope is to first generate the local explanations first and to then generate global level explanations by aggregating these. The main drawback in such approaches is the large runtime required to generate explanations for individual instances. Another way to address this problem is to create distilled models like decision trees for generating global explanations and local models for explanations at the instance level.

Lastly, trust is one of the most important aspects of interpretability. Consider the case of deep learning models that have shown great predictive performance in a number of healthcare applications [20]. While it is possible to extract explanations from deep learning models, these explanations cannot be proven to be sound or complete [23]. Often the goal of explanations is to get parsimonious explanations which cannot be stated to have the correct explanations. Additionally always having parsimony as a goal may lead to incorrect models [7].

D. Assistive Intelligence

One common misconception about the application of machine learning in healthcare is that machine learning algorithms are intended to replace human practitioners in healthcare and medicine [11]. Healthcare delivery is an extremely complex, subtle, and intimate process that requires domain knowledge and intervention in every step of care. We believe that the human healthcare practitioner will remain integral to their role and that machine learning algorithms can assist and augment the provision of better care. Human performance parity [17] is also considered to be an important aspect of predictive systems that provide explanations i.e., the predictive system should be at least as good as the humans in the domain and at least make the same mistakes that the human is making. In certain use cases the opposite requirement may hold i.e., one may not care about parity with cases when humans are right but rather one cares more about cases where humans are bad at prediction but the machine learning system has superior performance. Such hybrid of human-machine learning systems can lead to truly assistive machine Learning in healthcare. Explanations from such systems could also be used to improve human performance, extract insights, gain new knowledge which may be used to generate hypothesis etc. The results from hypothesis derived from the data driven paradigm could in turn be used to push the frontiers of knowledge in healthcare and medicine by guiding theory [2].

V. Interpretable Models in Healthcare

Depending upon the scope of the problem, explanations from machine learning models can be divided into “global”, “cohort-specific” and “local” explanations. Global explanations refer to explanations that apply to the entire study population e.g., in the case of decision trees and regression models. Cohort-specific explanations are explanations that are focal to population sub-groups. Local explanations refer to explanations that are at the instance level i.e., explanations that are generated for individuals. Consider Figure 2 which illustrates the contrast between global vs. local models for predicting diabetes. The global model is a decision tree model that generalizes over the entire population, the cohort level model can also be a decision tree model which captures certain nuances of the sub-population of patients not captured by the global model and lastly the local model gives explanations at the level of instances. All three explanations may be equally valid depending upon the use case and how much soundness and generalizability is required by the application.

One way to distinguish models is by model composition. The predictive model and the explanation of the model can be the same as in the case of decision trees, GA2M etc. Alternatively they can be different e.g., a Gradient Boosting model is not really interpretable but it is possible to extract explanations via models like LIME, Shapley values, Tree Explainers etc. One scheme to create interpretable models is via model distillation where the main idea is to create interpretable models from non-interpretable models. Consider a feature set $X = x_1, x_2, x_3, ..., x_n$ with y_i is the class label being predicted. Suppose y'_i is the label that is predicted by a prediction model M_p which is non-interpretable *e.g.,* Deep Learning etc. An interpretable model *e.g.,* decision trees, regression models etc. which is created by the feature set X and the output y'_i as the label is referred to as a student model. While there are no theoretical guarantees for the performance of the student model but in practice, many student models have predictive power which is sufficiently high from an application perspective.

VI. Future of Interpretability in Healthcare

As machine learning increasingly penetrates healthcare, issues around accountability, fairness and transparency of machine learning systems in healthcare will become paramount. Most predictive machine learning systems in healthcare just provide predictions but in practice many use cases do require reasoning to convince medical practitioners to take feedback from such models. Thus there is a need to integrate interpretable models with predictions with the workflow of medical facilities. Most predictive models are not prescriptive or causal in nature. In many healthcare applications explanations are not sufficient and prescriptions or actionability. We foresee causal explanations to be the next frontier of machine learning research.

It should also be noted that while interpretability is an aspect of holding machine learning models accountable, it is not the only way to do. Researchers have also suggested that one way to audit machine learning systems it to analyze their outputs given that some models may be too complex for human comprehension [29] and auditing outputs for fairness and bias may be a better option. Also, many problems in healthcare are complex and simplifying them to point solutions with accompanying explanations may result in suboptimal outcomes. Thus consider the problem of optimizing risk of readmission to a hospital. Just optimizing predictions and actionability to reducing risk of readmission may in fact increase the average length of stay in hospitals for patients. This would be non-optimal solution and not in the best interest of the patient even though the original formulation of the machine learning problem is defined as such. Thus problem formulations for interpretable models should take such contexts and interdependencies into account.

There is also some debate around the use of post-hoc vs. ante-hoc models of prediction in the research community. Since explanations from post-hoc models do not correspond to how the model actually predicts, there is skepticism regarding the use of these models in scenarios which may require critical decision making. Current and future efforts in predictive models should also focus on ante-hoc explanation models like context explanation networks, falling rule lists, SLIM etc. Scalability of interpretable machine learning models is also an open area of research. Generating explanations for models like LIME and Shapley values can be computationally expensive. In case of LIME, a local model has to be created for each instance for which an explanation is required. In a scenario where there are hundreds of millions of instances for which prediction and explanations are required then this can be problematic from a scalability perspective. Shapley values computation requires computing the variable contribution by considering all possible combinations of variables in the data. For problems where the feature set has hundreds of variables, such computations can be very expensive. The problem of scalability thus exists with two of the most widely used interpretable machine learning models.

Lastly, evaluation of explanation models is an area which has not been explored in much detail. Consider the scenario in which multiple models with the same generalization error offer different explanations for the same instance or alternatively different model agnostic models are used to extract explanations and these model offer different explanations. In both these scenarios, the challenge is to figure out which explanations are the best. We propose that the concordance in explanations as well as how well the explanations align with what is already known in the domain will determine explanation model preference. However, the danger also exists that novel but correct explanations may be weeded out if concordance is the only criteria of choosing explanations.

VII. Conclusion

Applied Machine Learning in Healthcare is an active area of research. The increasingly widespread applicability of machine learning models necessitates the need for explanations to hold machine learning models accountable. While there is not much agreement on the meaning of interpretability in machine learning, there are a number of characteristics of interpretable models that researchers have discussed which can be used as a guide to create the requirements of interpretable models. The choice of interpretable models depends upon the application an use case for which explanations are required. Thus a critical application like prediction a patient's end of life may have much more stringent conditions for explanation fidelity as compared to just predicting costs for a procedure where getting the prediction right is much more important as compared to providing explanations. There are still a large number of questions that are unaddressed in the area of interpretable models and we envision that it will be an active area of research for the next few years.

Acknowledgment

The authors would like to thank Kiyana Zolfaghar, Vikas Kumar, Dr. Nicholas Mark, Rohan D'Souza and other members of the KenSci team for the fruitful and enlightening discussions and feedback on interpretability in machine learning.

References

[1] TR Addis. Towards an" expert" diagnostic system. *ICL Technical Journal*, 1:79–105, 1956.

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

[3] Maruan Al-Shedivat, Avinava Dubey, and Eric P Xing. Contextual explanation networks. *arXiv preprint arXiv:1705.10301*, 2017.

[4] Jenna Burrell. How the machine thinks: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[7] Pedro Domingos. The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425, 1999.

[8] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[9] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

[10] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[11] Puneet Gupta. Machine learning: The future of healthcare. *Harvard Sci. Rev.*, 2017.

[12] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2125–2126. ACM, 2016.

[13] Alvin D Jeffery, Laurie L Novak, Betsy Kennedy, Mary S Dietrich, and Lorraine C Mion. Participatory design of probability-based decision support tools for in-hospital nurses. *Journal of the American Medical Informatics Association*, 24(6):1102–1110, 2017.

[14] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):15, 2012.

[15] Samantha Krening, Brent Harrison, Karen M Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55, 2017.

[16] Todd Kulesza. Personalizing machine learning systems with explanatory debugging. PhD Dissertation, Oregon State University, 2014.

[17] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[18] Gunasekaran Manogaran and Daphne Lopez. A survey of big data architectures and machine learning algorithms in healthcare. *International Journal of Biomedical Engineering and Technology*, 25(2-4):182–211, 2017.

[19] Tim Miller. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[20] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 2017.

[21] Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K Jha. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1893–1905, 2015.

[22] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

[23] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[24] Jesus Maria Pérez, Javier Muguerza, Olatz Arbelaitz, and Ibai Gurrutxaga. A new algorithm to build consolidated trees: study of the error rate and steadiness. In *Intelligent Information Processing and Web Mining*, pages 79–88. Springer, 2004.

[25] Marco Tulio Ribeiro, 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*, pages 1135–1144. ACM, 2016.

[26] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

[27] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Detecting bias in black-box models using transparent model distillation. *arXiv preprint arXiv:1710.06169*, 2017.

[28] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In *Companion of The Web Conference 2018*, pages 163–166. International World Wide Web Conferences Steering Committee, 2018.

[29] Michael Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, 2009.

[30] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

[31] Fulton Wang and Cynthia Rudin. Falling rule lists. In *Artificial Intelligence and Statistics*, pages 1013–1022, 2015.

[32] William Yang Wang. "Liar, Liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[33] Xin Yan and Xiaogang Su. *Linear Regression Analysis: Theory and Computing*. World Scientific, 2009.