In posting your proposed algorithm/article in this forum, it is required that you provide its detailed information including article abstract, article authors, article title, publication place (journal/conference names), publication year, page numbers, etc.

**Note: The subject of your message/post must be Option 1 followed by the title of the article you choose.**  
It will be helpful if you also post the article itself (pdf file or website URL) on Canvas. Also include your name and UCID in the proposal.

Finally, you MUST indicate the topic(s) that your proposed/algorithm/article is related to.

The topic(s) should be one of the following or a combination of the following topics:

**Deep Learning Algorithms and Architectures**

**Deep Learning Applications**

**Subject**: **Option 1 - Definitions, methods, and applications in interpretable machine learning**

**Topic**: Deep Learning Applications – interpretable machine learning

**Student**: Cunha, Wellington

**NJIT ID**: 31548454

**NJIT UC ID**: wc44

**Proposal**: discuss how interpretable machine learning can be applied into health-related fields, such as in algorithms for disease diagnosis. My starting article will be the main article below, but I have already found other articles that will be used in extending the discussions (all of the articles that I am currently working on are referenced below)

**Main Article**

Title: Definitions, methods, and applications in interpretable machine learning

Authors: Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B.

Publisher: Proceedings of the National Academy of Sciences

Date of Publication: July 1st, 2019

Pages: 22071–22080

Abstract

Machine-learning models have demonstrated great success in learning complex patterns that enable them to make predictions about unobserved data. In addition to using models for prediction, the ability to interpret what a model has learned is receiving an increasing amount of attention. However, this increased focus has led to considerable confusion about the notion of interpretability. In particular, it is unclear how the wide array of proposed interpretation methods are related and what common concepts can be used to evaluate them. We aim to address these concerns by defining interpretability in the context of machine learning and introducing the predictive, descriptive, relevant (PDR) framework for discussing interpretations. The PDR framework provides 3 overarching desiderata for evaluation: predictive accuracy, descriptive accuracy, and relevancy, with relevancy judged relative to a human audience. Moreover, to help manage the deluge of interpretation methods, we introduce a categorization of existing techniques into model-based and post hoc categories, with subgroups including sparsity, modularity, and simulatability. To demonstrate how practitioners can use the PDR framework to evaluate and understand interpretations, we provide numerous real-world examples. These examples highlight the often underappreciated role played by human audiences in discussions of interpretability. Finally, based on our framework, we discuss limitations of existing methods and directions for future work. We hope that this work will provide a common vocabulary that will make it easier for both practitioners and researchers to discuss and choose from the full range of interpretation methods.

Reference:

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

**References**

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018, August). Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics (pp. 559-560). <http://www.ieee-iib.org/2018/Aug/article1/iib_vol19no1_article1.pdf>

Karatekin, T., Sancak, S., Celik, G., Topcuoglu, S., Karatekin, G., Kirci, P., & Okatan, A. (2019, August). Interpretable machine learning in healthcare through generalized additive model with pairwise interactions (GA2M): predicting severe retinopathy of prematurity. In 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML) (pp. 61-66). IEEE. <https://ieeexplore.ieee.org/abstract/document/8876951>

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. <https://doi.org/10.1073/pnas.1900654116>

Okay, F. Y., Yıldırım, M., & Özdemir, S. (2021, October). Interpretable machine learning: a case study of healthcare. In 2021 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1-6). IEEE. <https://ieeexplore.ieee.org/abstract/document/9615727>

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning‐based prediction models in healthcare. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(5), e1379. <https://zitniklab.hms.harvard.edu/publications/papers/interpretableML-survey20.pdf>

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural computing and applications, 32(24), 18069-18083. <https://link.springer.com/article/10.1007/s00521-019-04051-w>