NEXT GENERATION CB HAZARD PREDICTION AND CONSEQUENCE ASSESSMENT WITH MULTI-ECHELON DECISION SUPPORT APPLICATIONS

Testing Methods For Interpreting Or Explaining Artificial Intelligence

James Lee NSWCIHD **Cody Youngbull** NSWCIHD

At present, systems that use artificial intelligence (AI) are being used and developed that have superior performance compared to traditional systems. This superior performance makes continued adoption almost inevitable. However, a large number of AI models and tools are black box systems. This is a major problem, how do you trust the predictions if you don't know how it's making decisions. This is especially true in life critical systems. A significant number of methods have been created to interpret or explain artificial intelligence systems. Some of these methods are specific to one type of system, like neural networks. Other methods claim to be agnostic and work across the range of machine learning methods. In terms of interpreting the functions within any black-box model, the LIME1 and SHAP2 methods are, by far, the most comprehensive and dominant across the literature methods for visualizing feature interactions and feature importance, while Friedman's PDPs,3 although much older and not as sophisticated, still remains a popular choice.4 It is critical to note that these methods were not developed by testing their performance on "non-black box" systems. This is an issue. How do you know "if you actually have a method of revealing what is in a black box" if that method is not tested on system where you know how it works (a transparent box or transparent neural network)? For this reason, we have constructed a transparent neural network (we know how it makes decisions) for testing these methods that claim to explain or interpret the behavior of AI systems. This is a critical capability that not only can be used to test current methods, but can aid in the development of new interpretation methods in the field of explainable artificial intelligence (XAI). This moves us toward being able to understand how AI systems make decisions so that we can trust them to aid in decision making.

Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
Friedman, J.H. Greedy function approximation: A gradient boosting machine. Ann. Stat. 2001, 29, 1189–1232.< Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods." Entropy 23, no. 1 (2020): 18.