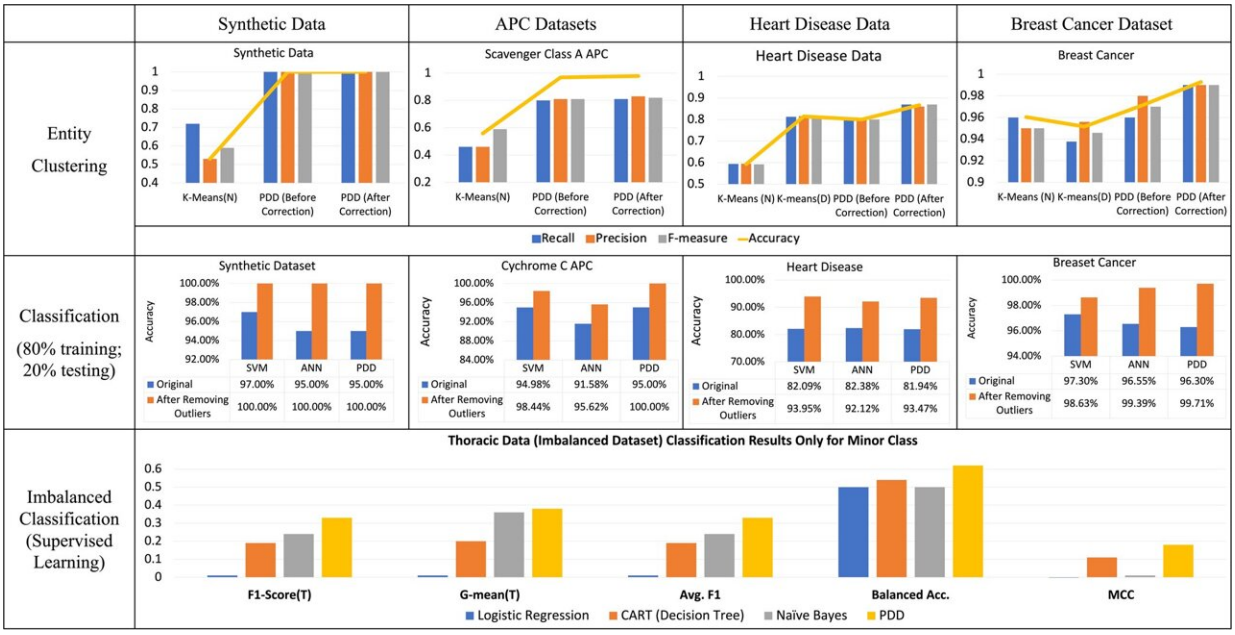
New model reduces bias and enhances trust in AI decision-making and knowledge organization

August 8 2023



Results Comparison of PDD and other ML Models . Credit: *npj Digital Medicine* (2023). DOI: 10.1038/s41746-023-00816-9

University of Waterloo researchers have developed a new explainable artificial intelligence (AI) model to reduce bias and enhance trust and accuracy in machine learning-generated decision-making and knowledge organization.

Traditional machine learning models often yield biased results, favoring groups with large populations or being influenced by unknown factors, and take extensive effort to identify from instances containing patterns and sub-patterns coming from different classes or primary sources.

The medical field is one area where there are severe implications for biased machine learning results. Hospital staff and medical professionals rely on datasets containing thousands of medical records and complex computer algorithms to make critical decisions about patient care.

Machine learning is used to sort the data, which saves time. However, specific patient groups with rare symptomatic patterns may go undetected, and mislabeled patients and anomalies could impact diagnostic outcomes. This inherent bias and pattern entanglement leads to misdiagnoses and inequitable health care outcomes for specific patient groups.

Thanks to new research led by Dr. Andrew Wong, a distinguished professor emeritus of systems design engineering at Waterloo, an innovative model aims to eliminate these barriers by untangling complex patterns from data to relate them to specific underlying causes unaffected by anomalies and mislabeled instances. It can enhance trust and reliability in Explainable Artificial Intelligence (XAI.)

The study, "Theory and rationale of interpretable all-in-one pattern discovery and disentanglement system," appears in the journal *npj Digital Medicine*.

"This research represents a significant contribution to the field of XAI," Wong said. "While analyzing a vast amount of protein binding data from X-ray crystallography, my team revealed the statistics of the physicochemical amino acid interacting patterns which were masked and mixed at the data level due to the entanglement of multiple factors

present in the binding environment. That was the first time we showed entangled statistics can be disentangled to give a correct picture of the deep knowledge missed at the data level with scientific evidence."

This revelation led Wong and his team to develop the new XAI model called Pattern Discovery and Disentanglement (PDD).

"With PDD, we aim to bridge the gap between AI technology and human understanding to help enable trustworthy decision-making and unlock deeper knowledge from complex data sources," said Dr. Peiyuan Zhou, the lead researcher on Wong's team.

Professor Annie Lee, a co-author and collaborator from the University of Toronto, specializing in natural language processing, foresees the immense value of PDD contribution to clinical decision-making.

The PDD model has revolutionized pattern discovery. Various case studies have showcased PDD, demonstrating an ability to predict patients' medical results based on their clinical records. The PDD system can also discover new and rare patterns in datasets. This allows researchers and practitioners alike to detect mislabels or anomalies in machine learning.

The result shows that health care professionals can make more reliable diagnoses supported by rigorous statistics and explainable patterns for better treatment recommendations for various diseases at different stages.

More information: Andrew K. C. Wong et al, Theory and rationale of interpretable all-in-one pattern discovery and disentanglement system, *npj Digital Medicine* (2023). DOI: 10.1038/s41746-023-00816-9

Provided by University of Waterloo

Citation: New model reduces bias and enhances trust in AI decision-making and knowledge organization (2023, August 8) retrieved 9 May 2024 from https://techxplore.com/news/2023-08-bias-ai-decision-making-knowledge.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.