DEPARTMENT OF COMPUTER SCIENCE

TEXAS TECH
Whitacre College of Engineering

Concept-based Interpretability for Responsible AI

Aidong Zhang, Ph.D.
University of Virginia
Tuesday, October 21, 2025
3:30 p.m.
Zoom

Abstract: In recent years, major advances in artificial intelligence (AI) have been applied to medical image diagnosis with promising results. Even though these methods demonstrate incredible potential in saving valuable man-hours and minimizing inadvertent human mistakes, their adoption has been met with rightful skepticism and extreme circumspection in critical applications such as medical diagnosis. The most paramount of these challenges is the lack of rationale behind predictions - making them notoriously a black box in nature. In extreme cases, this can create a lack of alignment between the designer's intended behavior and the model's actual performance. In this talk, I will discuss our recent research on explainable AI strategies, in particular, I will discuss concept-based learning models and show how the concept-based learning models and example-based learning models can be designed for explainable deep neural networks, vision transformers, and vision language models.

Bio: Dr. Aidong Zhang is Thomas M. Linville Endowed Professor of Computer Science in the School of Engineering and Applied Sciences at University of Virginia (UVA). She also holds joint appointments with Department of Biomedical Engineering and School of Data Science at University of Virginia. Her research interests include machine learning, data mining, bioinformatics, and health informatics. Dr. Zhang is a fellow of ACM (Association for Computing Machinery), AIMBE (American Institute for Medical and Biological Engineering), and IEEE (Institute of Electrical and Electronics Engineers). She is also a member of the Virginia Academy of Science, Engineering and Medicine (VASEM).