Thesis Defense

Computer Science Master's Program

"Ensemble Graph Attention Networks Enable Interpretable Biomarker Discovery for Breast Cancer Subtyping"

By **Nicholas Zarate**

Abstract:

The reproducibility crisis in biomarker discovery stems from traditional approaches that often treat molecular features as independent variables while ignoring the networked nature of biological systems. We present an interpretable-by-design framework that models individual tumor network states using graph attention networks (GATs) to discover robust breast cancer biomarkers. By constraining the search space through biologically-informed gene selection and multi-relational graphs integrating protein-protein interactions, pathways, and co-expression networks, we guide the model toward genuine biological relationships rather than spurious correlations. Our ensemble GAT approach achieved 76.8% balanced accuracy for molecular subtype classification. Systematic analysis of attention weights revealed an unexpected finding: 95 of 96 high-confidence biomarkers were terminal nodes rather than network hubs, consistently connecting to established breast cancer drivers including TP53, EGFR, ESR1, and CCND1. We successfully distilled these network-based discoveries into an 85-gene diagnostic panel using interpretable linear models, achieving 74.8% accuracy with expression data alone. Our biologically-constrained, interpretable-by-design approach demonstrates how network-guided machine learning yields both mechanistic understanding and reproducible biomarkers.

Date: Tuesday, November 18th, 2025
Time: 12:00 PM – 2:00 PM
Location: 14-238b
Committee: Dr. Anderson, Dr. Davidson, and Dr. Migler