A Study on Predicting Cancer Therapy-Related Cardiac Dysfunction in Cardio-Oncology Patients Using a Machine Learning Approach

Cheongin Rachel Im, PhD, MSN, RN[1,2], Julie A. Zuniga, PhD, RN, FAAN[2]
[1]UT Health San Antonio School of Nursing, [2]University of Texas at Austin School of Nursing

The University of Texas at Austin School of Nursing
UT Health San Antonio
All of Us RESEARCH PROGRAM
AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.

Background

- Cardio-oncology patients face overlapping symptoms and complex treatment challenges, which may increase their risk of developing cancer therapy-related cardiac dysfunction (CTRCD), including heart failure and cardiomyopathy.
- Identifying key factors that predict the occurrence of CTRCD is crucial for improving prognosis.

Objectives

This study aims to develop an ML-based predictive model for CTRCD in cardio-oncology patients and to identify key predictors that contribute to model performance.

Research Design Methodology

- This study used All of Us Research Program data, incorporating Electronic Health Record and survey responses.
- The final sample included 5,729 participants, with 19% having CTRCD and 81% without CTRCD.
- To address class imbalance, class weighting was applied.
- We implemented three Machine Learning models: Logistic Regression, Random Forest, and XGBoost.
- For data preprocessing, the dataset was split into Train – Test (80/20) subsets, and robust scaling and one-hot encoding were applied.
- Each model trained using the training dataset and evaluated based on predictions made on the test dataset to access performance.

Acknowledgements

This study was part of my doctoral dissertation. I would like to thank my committee members (Dr. Kavita Radhakrishnan, Dr. Heather Becker, and Dr. Yan Liu) for their valuable advice on this research, as well as Dr. Shelli Kesler for her statistical advice. I would also like to thank the University of Texas at Austin for their support of my research, and my current Principal Investigator, Dr. Lixin Song.

Results

- Among the three machine learning models (Figure 1), XGBoost demonstrated the most balanced performance, achieving high accuracy (0.815) and an F1 score (0.896).
- The AUC was 0.702, indicating fair to good predictive ability.
- Feature importance analysis revealed that the total number of symptoms was the strongest feature in the predictive model on CTRCD, followed by age, income, gender, and key social determinants of health (SDOH) such as neighborhood disorder and discrimination (Figure 2).

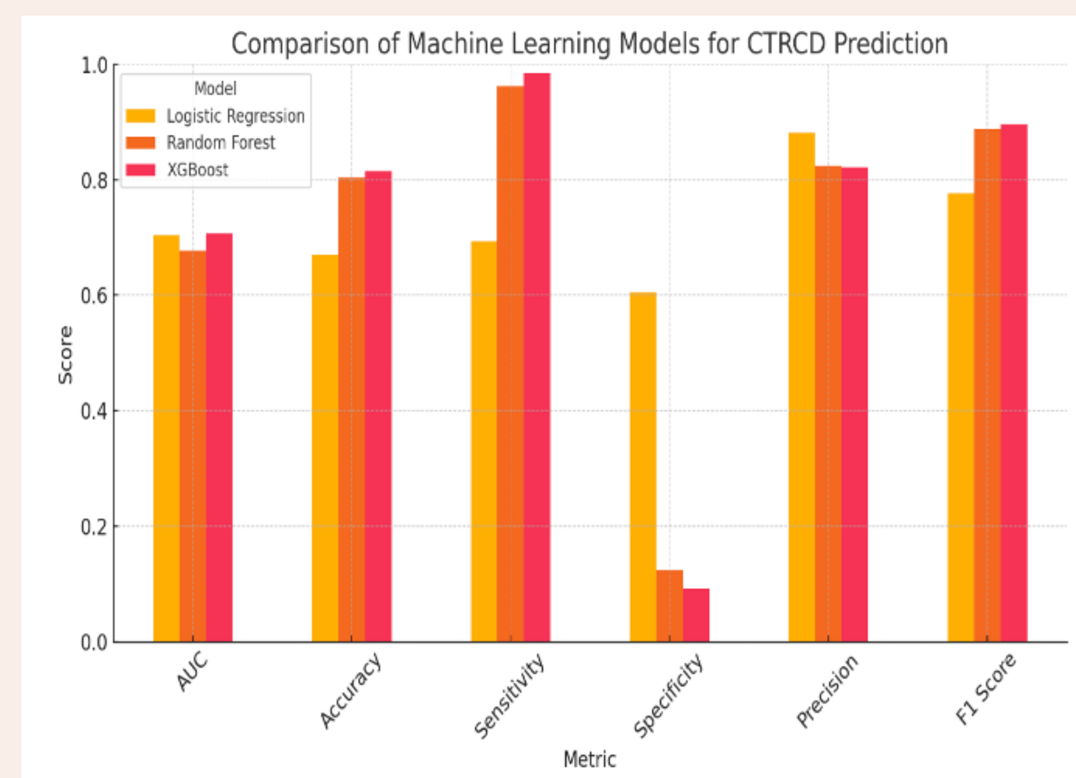

Figure 1. Model Comparison

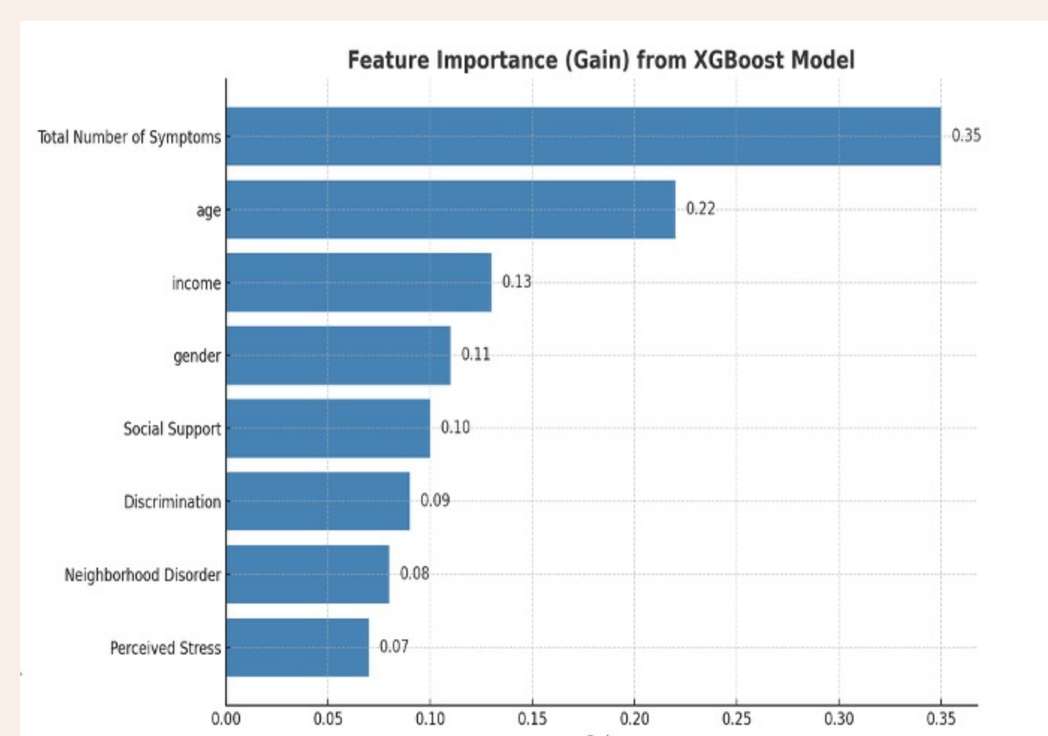

Figure 2. Feature Importance (Gain)

Conclusion

XGBoost effectively predicted CTRCD risk in cardio-oncology patients, with symptom burden and SDOH emerging as key predictors. Integrating machine learning-based tools into clinical care may support early risk detection and guide patient-centered interventions.

Future Plans

Future research will overcome the current study's limitations, such as limited specificity and reliance on secondary data, by establishing a balanced comparison group and employing standardized instruments to assess symptoms and CTRCD through primary data collection.