COLLEGE OF ENGINEERING AND COMPUTER SCIENCE
FLORIDA ATLANTIC UNIVERSITY

Announces the Ph.D. Dissertation Defense of

# Aaron N. Richter

for the degree of Doctor of Philosophy (Ph.D.)

## "Predicting Melanoma Risk from Electronic Health Records with Machine Learning Techniques"

July 12, 2019, 10:00 a.m.
Engineering East, Room 405
777 Glades Road
Boca Raton, FL

DEPARTMENT:
Computer and Electrical Engineering and Computer Science

ADVISOR:
Taghi M. Khoshgoftaar, Ph.D.

PH.D. SUPERVISORY COMMITTEE:

Taghi M. Khoshgoftaar, Ph.D., Chair
Mehrdad Nojoumian, Ph.D.
Dingding Wang, Ph.D.
Xingquan Zhu, Ph.D.

ABSTRACT OF DISSERTATION
Machine Learning Techniques for Predicting Melanoma Risk from Electronic Health Records

Melanoma is one of the fastest growing cancers in the world, and can affect patients earlier in life than most other cancers. Therefore, it is imperative to be able to identify patients at high risk for melanoma and enroll them in screening programs to detect the cancer early. Electronic health records collect an enormous amount of data about real-world patient encounters, treatments, and outcomes. This data can be mined to increase our understanding of melanoma as well as build personalized models to predict risk of developing the cancer. Cancer risk models built from structured clinical data are limited in current research, with most studies involving just a few variables from institutional databases or registries. This dissertation presents data processing and machine learning approaches to build melanoma risk models from a large database of de-identified electronic health records. The database contains consistently captured structured data, enabling the extraction of hundreds of thousands of data points each from millions of patient records. Several experiments are performed to build effective models, particularly to predict sentinel lymph node metastasis in known melanoma patients and to predict individual risk of developing melanoma. Data for these models suffer from high dimensionality and class imbalance. Thus, classifiers such as logistic regression, support vector machines, random forest, and XGBoost are combined with advanced modeling techniques such as feature selection and data sampling. Risk factors are evaluated using regression model weights and decision trees, while personalized predictions are provided through random forest decomposition and Shapley additive explanations.

Random undersampling on the melanoma risk dataset shows that many majority samples can be removed without a decrease in model performance. To determine how much data is truly needed, we explore learning curve approximation methods on the melanoma data and three publicly-available large-scale biomedical datasets. We apply an inverse power law model as well as introduce a novel semi-supervised curve creation method that utilizes a small amount of labeled data.

BIOGRAPHICAL SKETCH
Born in Boca Raton, Florida
B.S. 2014, Florida Atlantic University, Boca Raton, Florida
M.S. 2018, Florida Atlantic University, Boca Raton, Florida
Ph.D. 2019, Florida Atlantic University, Boca Raton, Florida

CONCERNING PERIOD OF PREPARATION
& QUALIFYING EXAMINATION

**Time in Preparation:** 2015 - 2019

**Qualifying Examination Passed:** Fall 2014
**Selected Published Papers:**

A. N. Richter and T. M. Khoshgoftaar, "Efficient learning from big data for cancer risk modeling: A case study with melanoma," *International Journal of Computers in Biology and Medicine*, vol. 110, pp. 29–39, Jul. 2019.

A. N. Richter and T. M. Khoshgoftaar, "Melanoma risk modeling from limited positive samples," *International Journal of Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 8, pp. 7, Dec. 2019.

A. N. Richter and T. M. Khoshgoftaar, "A review of statistical and machine learning methods for modeling cancer risk using structured clinical data," *International Journal of Artificial Intelligence in Medicine*, vol. 90, pp. 1–14, Aug. 2018.

A. N. Richter and T. M. Khoshgoftaar, "Melanoma Risk Prediction with Structured Electronic Health Records," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18*, Washington, DC, USA, 2018, pp. 194–199.

A. N. Richter and T. M. Khoshgoftaar, "Predicting sentinel node status in melanoma from a real-world EHR dataset," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Kansas City, MO, 2017, pp. 1872–1878.