COLLEGE OF ENGINEERING AND COMPUTER SCIENCE
FLORIDA ATLANTIC UNIVERSITY

Announces the Ph.D. Dissertation Defense of

John T. Hancock III

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

"Fraud Detection in Highly Imbalanced Big Data with Novel and Efficient Data Reduction Techniques"

April 8, 2024, 10:30 a.m.
In-person Room EE-405

DEPARTMENT:
Electrical Engineering and Computer Science

ADVISOR:
Taghi M. Khoshgoftaar, Ph.D.

Ph.D. SUPERVISORY COMMITTEE:
Taghi M. Khoshgoftaar, Ph.D., Chair
Man Chon U, Ph.D.
DingDing Wang, Ph.D.
Mehrdad Nojoumian, Ph.D.

ABSTRACT OF DISSERTATION

The rapid growth of digital transactions and the increasing sophistication of fraudulent activities have necessitated the development of robust and efficient fraud detection techniques, particularly in the financial and healthcare sectors. This dissertation focuses on the use of novel data reduction techniques for addressing the unique challenges associated with detecting fraud in highly imbalanced Big Data, with a specific emphasis on credit card transactions and Medicare claims. The highly imbalanced nature of these datasets, where fraudulent instances constitute less than one percent of the data, poses significant challenges for traditional machine learning algorithms. This dissertation explores novel data reduction techniques tailored for fraud detection in highly imbalanced Big Data. The primary objectives include developing efficient data preprocessing and feature selection methods to reduce data dimensionality while preserving the most informative features, investigating various machine learning algorithms for their effectiveness in handling imbalanced data, and evaluating the proposed techniques on real-world credit card and Medicare fraud datasets.

This dissertation covers a comprehensive examination of datasets, learners, experimental methodology, sampling techniques, feature selection techniques, and hybrid techniques. Key contributions include the analysis of performance metrics in the context of newly available Big Medicare Data, experiments using Big Medicare data, application of a novel ensemble supervised feature selection technique, and the combined application of data sampling and feature selection. The research demonstrates that, across both domains, the combined application of random undersampling and ensemble feature selection significantly improves classification performance.

The contributions presented advance the field of fraud detection. They add to the development of fraud detection systems. The proposed data reduction techniques offer practical solutions for tackling the challenges associated with imbalanced and high-dimensional data, offering a means to reduced financial losses, improved patient care, and increased trust in critical financial and healthcare domains. This dissertation offers practical implications for the financial and healthcare sectors by improving the detection of fraudulent transactions and Medicare Insurance claims. The proposed techniques may also be considered for general machine learning applications.

BIOGRAPHICAL SKETCH

Born in Florida, USA
B.S., University of Maryland University College 2009
M.S., Florida Atlantic University, Boca Raton, Florida 2015
Ph.D., Florida Atlantic University, Boca Raton, Florida 2024

CONCERNING PERIOD OF PREPARATION
& QUALIFYING EXAMINATION

Time in Preparation: 2019- 2024

Qualifying Examination Passed: Spring 2019

Published Papers:

J. T. Hancock and T. M. Khoshgoftaar. Catboost for big data: an interdisciplinary review. Journal of big data, 7(1):1–45, 2020.

J. T. Hancock and T. M. Khoshgoftaar. Medicare fraud detection using catboost. In 2020 IEEE 21st international conference on information reuse and integration for data science (IRI), pages 97–103. IEEE, 2020

J. T. Hancock and T. M. Khoshgoftaar. Performance of catboost and xgboost in medicare fraud detection. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 572–579. IEEE, 2020.

J. T. Hancock and T. M. Khoshgoftaar. Gradient boosted decision tree algorithms for medicare fraud detection. Springer Nature Computer Science Journal, 2(4):1–12, 2021.

J. T. Hancock and T. M. Khoshgoftaar. Impact of hyperparameter tuning in classifying highly imbalanced big data. In 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), pages 348–354. IEEE, 2021

J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson. Exploring area under the precision recall curve and random undersampling to classify imbalanced big data. Proceedings of the 27th ISSAT International Conference on Reliability and Quality in Design, page 111–116, 2022.

J. T. Hancock and T. M. Khoshgoftaar. Optimizing ensemble trees for big data healthcare fraud detection. In 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), pages 243–249. IEEE, 2022

J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson. A comparative approach to threshold optimization for classifying imbalanced data. In The International Conference on Collaboration and Internet Computing (CIC). IEEE, 2022.

J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson. The effects of random undersampling for big data medicare fraud detection. In 2022 IEEE International Conference on Service-Oriented System Engineering (SOSE), 2022. IEEE, 2022.

J. Hancock and T. M. Khoshgoftaar. Data reduction to improve the performance of one-class classifiers on highly imbalanced big data. In 2023 22nd IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2023.

J. T. Hancock, R. A. Bauder, and T. M. Khoshgoftaar. A model-agnostic feature selection technique to improve the performance of one-class classifiers. In 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2023

J. T. Hancock and T. M. Khoshgoftaar. Exploring maximum tree depth and random undersampling in ensemble trees to optimize the classification of imbalanced big data. Springer Nautre Computer Science Journal, 4(5):462, 2023

J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar. Explainable machine learning models for medicare fraud detection. Journal of Big Data, 10(1):154, 2023.

J. T. Hancock, H. Wang, T. M. Khoshgoftaar, and Q. Liang. Data reduction techniques for highly imbalanced medicare big data. Journal of Big Data, 11(1):8, 2024.