FAU COLLEGE OF ENGINEERING
AND COMPUTER SCIENCE
FLORIDA ATLANTIC UNIVERSITY

Announces the Ph.D. Dissertation Defense of

Mary Anne Walauskis

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

“Novel Unsupervised Frameworks for Automated Class Distribution Estimation and Binary Label Generation for Tabular Data”

November 3, 2025 at 10:30 a.m.
Virtual Presentation: https://fau-edu.zoom.us/j/89820447209?pwd=2JURrrnbpIx1chcnf0DVOHCaHUBgsX.1
Meeting ID: 898 2044 7209 Passcode: rNh5PT

DEPARTMENT:
Electrical Engineering and Computer Science

ADVISOR:
Taghi M. Khoshgoftaar, Ph.D.

Ph.D. SUPERVISORY COMMITTEE:
Taghi M. Khoshgoftaar, Ph.D., Chair
Mohammad Ilyas, Ph.D.
Mehrdad Nojoumian, Ph.D.
DingDing Wang, Ph.D.

ABSTRACT OF DISSERTATION

This data has the potential to accelerate machine learning research; however, supervised methods require labels, and unsupervised methods often require expert fine-tuning to be reliable, both of which can impose significant cost. In addition to not requiring labels, another benefit of unsupervised learning is the protection of privacy since it does not require human annotation. In addition, class imbalance, where one class has significantly more instances, can complicate model training and reduce performance. Because of these challenges, automated unsupervised methods can offer a path forward to further machine learning research. The primary objective of this dissertation is to develop a novel method for determining the class distribution of an unlabeled dataset, along with a fully automated and unsupervised class labeling framework. We validate our methods across a diverse set of real-world tabular datasets that vary widely in domain, class distribution, feature dimensionality, and size, including challenging applications such as fraud detection and cognitive assessment.

Our unique approach involves the combination of two labeling strategies, an unsupervised ensemble and percentile-threshold based methods, that create a high-confidence set of labels which ultimately determine a single positive or negative label for each instance in the dataset based on the expected number of positives. We further improve label quality and efficiency by integrating unsupervised feature selection to rank and identify the most informative features. Unsupervised feature selection simplifies the model and reduces computational complexity, making the method well-suited for large-scale, severely imbalanced datasets (e.g., Medicare and credit-card fraud). Moreover, we enhance our labeling method by introducing an unsupervised framework that automatically estimates the class distribution. Using this estimate, the framework selects decision thresholds adaptively, thereby improving label quality. Our novel approach relies exclusively on the dataset’s own features for labeling, requiring no external labels or manual annotations. This makes the method fully automated and unsupervised. We detail empirical results demonstrating substantial improvements in label quality, both across refinements of the method (e.g., progressing from unsupervised to automated unsupervised approaches) and in comparison to an unsupervised baseline learner. These results highlight the effectiveness of our novel class distribution estimation and class label generation methods when applied to unlabeled data.

BIOGRAPHICAL SKETCH

Born in Maryland, USA
B.S., Valencia College, Orlando, Florida 2021
M.S., University of West Florida, Pensacola, Florida 2022
Ph.D., Florida Atlantic University, Boca Raton, Florida 2025

CONCERNING PERIOD OF PREPARATION
& QUALIFYING EXAMINATION

Time in Preparation: 2023-2025

Qualifying Examination Passed: Fall 2023

Published Papers:

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "Confident labels: A novel approach to new class labeling and evaluation on highly imbalanced data." *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2024. Won the Best Student Paper Award.

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "New class labeling and evaluation methodology for balanced and highly imbalanced data." *2024 International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2024.

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "Unsupervised label generation for severely imbalanced fraud data." *Journal of Big Data* 12.1 (2025): 63.

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "SHAP-based Feature Selection for Enhanced Unsupervised Labeling." *IEEE Access* 99 (2025): 1-1.

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "An Automated Framework for Unsupervised Binary Class Distribution Estimation." *2025 IEEE 11th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*. IEEE, 2025.

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "Choosing the right metrics: A study of performance measurement for binary classification in imbalanced and big data." *The International FLAIRS Conference Proceedings*. Vol. 38. No. 1. 2025.

Hancock, John T., Robert K.L. Kennedy, Mary Anne Walauskis, and Taghi M. Khoshgoftaar. "A New and Effective Technique for Unsupervised Labeling and Feature Selection with Applications in Healthcare Fraud Detection." *2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI)*. IEEE, 2025.

Salekshahrezaee, Zahra, Mary Anne Walauskis, and Taghi M. Khoshgoftaar. "Unsupervised Feature Extraction using Convolutional Autoencoder for Credit Card Fraud Detection." *2025 International Conference on Machine Learning and Applications (ICMLA)*. (Accepted, Forthcoming 2025).

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "Scalable Unsupervised Labeling with SHAP Feature Selection for Fraud Detection in Imbalanced Data." *Journal of Big Data* (Accepted, Forthcoming 2025).

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "A Novel Approach to Automating Unsupervised Estimation of Class Distribution." *Journal of Big Data* (Accepted, Forthcoming 2025).

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "A Study of Ensemble-Gradient and Autoencoder Binary Labeling in Imbalanced Fraud Datasets." *Journal of Big Data* (Submitted, Under review 2025).

Walauskis, Mary Anne, and Taghi M. Khoshgoftaar. "Labeling Without Limits: An Automated Unsupervised Labeling Approach for Tabular Data." *IEEE Access*, IEEE (Submitted, Under review 2025).