COLLEGE OF ENGINEERING AND COMPUTER SCIENCE
FLORIDA ATLANTIC UNIVERSITY

Announces the Ph.D. Dissertation Defense of

Tawfiq Hasanin

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

“Investigating Machine Learning Algorithms with Imbalanced Big Data”

July 1, 2019, 10:30 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
Hanqi Zhuang, Ph.D.
Bassem Alhalabi, Ph.D.
Xingquan Zhu, Ph.D.

ABSTRACT OF DISSERTATION

Recent technological developments have engendered an expeditious production of big data and also enabled machine learning algorithms to produce high-performance models from such data. Nonetheless, class imbalance (in binary classifications) between the majority and minority classes in big data can skew the predictive performance of the classification algorithms toward the majority (negative) class whereas the minority (positive) class usually holds greater value for the decision makers. Such bias may lead to adverse consequences, some of them even life-threatening, when the existence of false negatives is generally costlier than false positives. The size of the minority class can vary from fair to extraordinary small, which can lead to different performance scores for machine learning algorithms. Class imbalance is a well-studied area for traditional data, i.e., not big data. However, there is limited research focusing on both rarity and severe class imbalance in big data. This dissertation subsumes nine case studies, utilizing three learners, six data sampling approaches, one feature selection, three performance metrics, three model evaluation strategies, and various class distribution ratios, to uniquely investigate the effect of machine learning with class imbalance levels (high, severe, and rare) in big data analytics. Model performance varies depending on the characteristics of the data. We show, for the most part, that undersampling outperforms oversampling where undersampling imposes a lower computational burden and results in a faster model training time, which is beneficial to big data analytics. Moreover, we show that forcing the majority class to become the minority, may lead to algorithm performance improvement. One of our proposed solutions, a constitution of Random Undersampling and Feature Importance, had a higher prediction performance compared to that of the highest value of the winning algorithm of an international bioinformatics machine learning competition. Additionally, we artificially injected rarity to study its impact on model performance. Finally, we artificially injected an imbalanced positive class condition into big balanced data, in which our work shows that model performance across imbalanced big data can be effectively improved using undersampling, without the need to considerably alter the composition of the original data and reach the perfectly balanced size.

BIOGRAPHICAL SKETCH
Born in Mecca, Saudi Arabia
B.S., King Abdulaziz University, Jeddah, Saudi Arabia, 2001
M.S., King Abdulaziz University, Jeddah, Saudi Arabia, 2009
Ph.D., Florida Atlantic University, Boca Raton, Florida, 2019

CONCERNING PERIOD OF PREPARATION
& QUALIFYING EXAMINATION

Time in Preparation: 2014-2019

Qualifying Examination Passed: Semester Spring 2014

Selected Published Papers:

Tawfiq Hasanin, Taghi M Khoshgoftaar, Joffrey L Leevy, and Naeem Seliya. Examining characteristics of predictive models with imbalanced big data. Journal of Big Data, 20 pages, 2019 (Under Review).

Tawfiq Hasanin, Taghi M Khoshgoftaar, and Richard A Bauder. Chapter: Experimental studies on the impact of data sampling with severely imbalanced big data. In Reuse in intelligent systems, Cambridge Press, 21 pages, 2019 (Accepted).

Tawfiq Hasanin and Taghi M Khoshgoftaar. The effects of random undersampling with simulated class imbalance for big data. In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pages 70–79. IEEE, 2018.

Richard A Bauder, Taghi M Khoshgoftaar, and Tawfiq Hasanin. Data sampling approaches with severely imbalanced big data for Medicare fraud detection. In 2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI), pages 137–142. IEEE, 2018.

Sara Landset, Taghi M Khoshgoftaar, Aaron N Richter, and Tawfiq Hasanin. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, 2(1):24, 36 pages, 2015.