



Accessible Data Mining



Department of Computer Science

University of Toronto, Toronto, ON M5S 2E4, Canada



My objective is to pursue a Computer Science Ph.D. with a focus on **data mining (DM) algorithms, systems, and applications**. I received my degrees from University of Cincinnati (BSc in computer engineering) and University of Toronto (MSc in computer science) in 2015 and 2017. Since graduation, I have been working at PwC Canada, a leading global consulting firm, as an applied machine learning scientist. After obtaining my Ph.D. from KCL, I would like to continue with a research career, and bridge the gap between academia and industry by using emerging DM models to solve real-world problems.

My past research has yielded academic outcomes in three categories: (i) fundamental DM algorithms [2, 3, 7]; (ii) scalable learning systems [8] and (iii) interdisciplinary studies between DM and other fields [1, 4, 5, 6]. I have extensive experience in outlier detection (anomaly detection), a key area for identifying anomalous data objects from the general data distribution, with numerous use cases like fraudulent financial activity prevention and network intrusion detection. In 2018, I designed and implemented `Python Outlier Detection Toolbox (PyOD)`; it quickly became the most popular open-source detection library and made top 20 DM libraries on GitHub¹. Being widely used in both academic and commercial projects, PyOD receives 20,000 downloads and 600 GitHub stars in six months. An accompanying paper is being reviewed at *The Journal of Machine Learning Research* [8].

Firstly, I have proposed a series of **semi-supervised and unsupervised algorithms** to facilitate reliable outlier detection with limited data using diversified instruments [2, 3, 7]. One of the models, `XGBOD`, augments the original feature space by unsupervised feature engineering for richer data representation; its semi-supervised ensembling approach leads to superior detection ability on imbalanced outlier datasets [3]. Another model, `DCSO`, emphasizing the importance of selecting competent base detectors locally during model combination [2], was presented at *2018 KDD conference*. I then proposed `LSPC` based on `DCSO` with refined schemes for detector selection, which is recently accepted at *SDM 2019* [7].

Secondly, the absence of **comprehensive and scalable learning systems** prevents researchers from conducting fair model comparisons and practitioners from doing agile developments. This motivates me to create the learning frameworks like `PyOD`; the toolkit is not only acknowledged for its comprehensiveness with more than 20 detection algorithms included (e.g., neural network-based models and outlier ensembles), but also for its optimized model efficiency by just-in-time compilation (JIT) and parallelization. In late 2018, I initialized `Julia Outlier Detection Library (OutlierDetection.jl)` to provide multi-

¹<https://github.com/yzhao062/pyod>

language support for broader audiences. Notably, I was working at Siemens from 2012–2014 as a C++ engineer to develop and maintain **Siemens NX**, one of the most widely used CAD software. This experience, along with multiple programming teaching assistantships, ensures my competency in low-level programming and scalable system design.

Thirdly, applying DM techniques to practical problems can be challenging but also rewarding, which encourages me to conduct **interdisciplinary research between DM and other fields** such as music [1], medical [4], security [5], and human resource management [6]. In our paper presented at *Ubicomp/ISWC 2017*, we evaluated the performances of existing smartwatch authentication systems and proposed a recommender system approach as a future direction for robustness enhancement [5]. In a recent work of analyzing employee attrition, we conducted turnover forecasting with the latest predictive models, analyzed the results by statistical tests, and proposed a reliable way of turnover prediction [6].

Future Research Plan

I am applying to the Ph.D. program at KCL due to its leading position in data mining. I find the work of **Prof. Lorenzo Cavallaro** particularly relevant to my interests, i.e., leveraging DM techniques in system security research. Although my rich experience in **anomaly detection** inspires me to focus on this field, I am open to other emerging interdisciplinary directions. My ultimate goal is to eliminate the barriers between people and technology and make understanding and applying DM methods more accessible.

The first research goal is to design *near unsupervised learning algorithms*, under the constraint of the limited number of ground truth. Notably, ensemble learning can be useful to control model variance, and I am interested in combining ensemble learning with representation learning and active learning together. For instance, one could construct a set of unsupervised base models to generate pseudo ground truth; a pseudo-supervised model can then be “trained” unsupervisedly. This can be further enhanced with active learning; multi-level knowledge, e.g., the root cause of an anomaly, can be gathered to offer a more holistic view during human verification.

The second goal is to develop scalable and easy-to-use learning systems by marrying my knowledge in programming languages, system design, and machine learning. Besides the JIT and parallelization used in PyOD, more considerations can be given to task-oriented systems. For instance, an appropriate data structure could speed up the model execution, e.g., using k-d trees in nearest neighbors algorithms. Certain tasks like subspace algorithms can be good cases to be implemented as distributed systems.

The third goal is to conduct interdisciplinary studies to make DM methods more approachable. One promising area is the intersection of security and DM. For instance, anomaly detection could be used to replace existing rule-based models in malware detection (treating malware as anomalies). The hybrid approach, using both predictive models and existing rules, may yield significant performance improvement. Compared with traditional rule-based systems, it can also catch unseen patterns when concept drift happens.