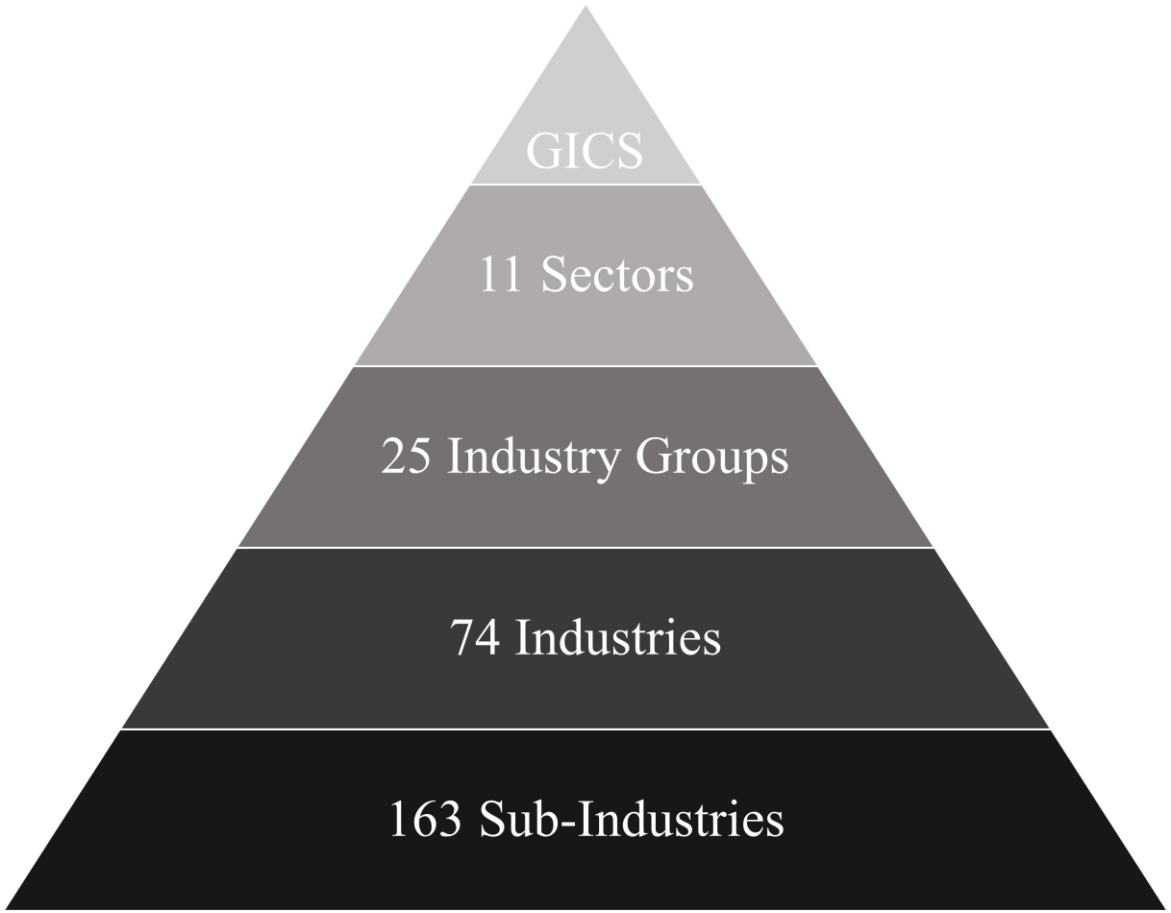
University of St Andrews

Master's Dissertation

Industry Classification Prediction Using Hierarchical Classification

Ziteng Dong **Supervisor: Dr Giorgos Minas**

School of Mathematics and Statistics



Hierarchical Structure of GICS

Research Target:
Predicting Global Industry Classification Standard (GICS)

Facts about GICS:

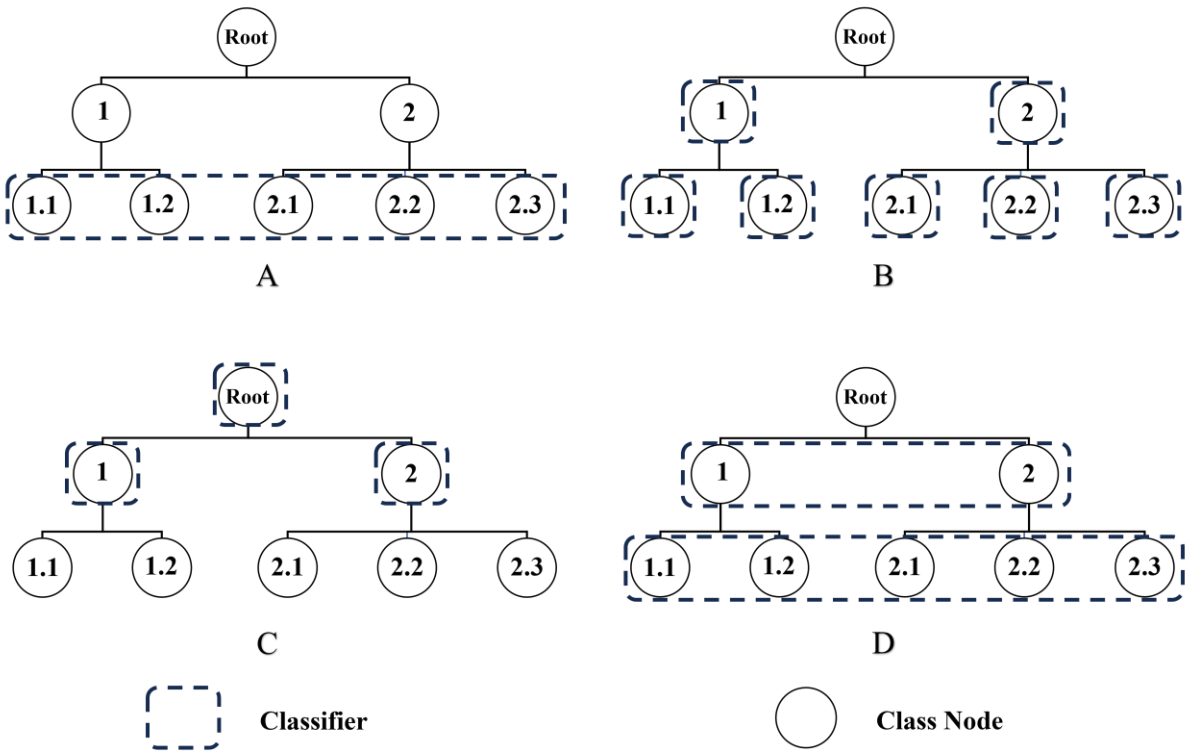
- A classification system for listed companies around the world.
- 4 Levels and 8 digits
- Example of Microsoft

Hierarchy	Codes	Text description
Sector	45	Information Technology
Industry Group	4510	Software & Services
Industry	451030	Software
Sub-Industry	45103020	Systems Software

Research Method:
Hierarchical classification algorithm (on Python)

Result:

- Successfully predict GICS codes with high accuracy using accounting data;
- Demonstrate the superiority of hierarchical classifications compared with flat ones



Visual Representation of Flat and Local Classifier Approaches

A: Flat approach
B: Local classifier per node (most popular and my choice)
C: Local classifier per parent node
D: Local classifier per level

Difference between *hierarchical* and *traditional* (*flat*) classification tasks

- Existence of local classifier
- Ability to handle information hidden in the taxonomy
- Definition of right and wrong classifications
  (*The 'why did I choose this?' moment!)*

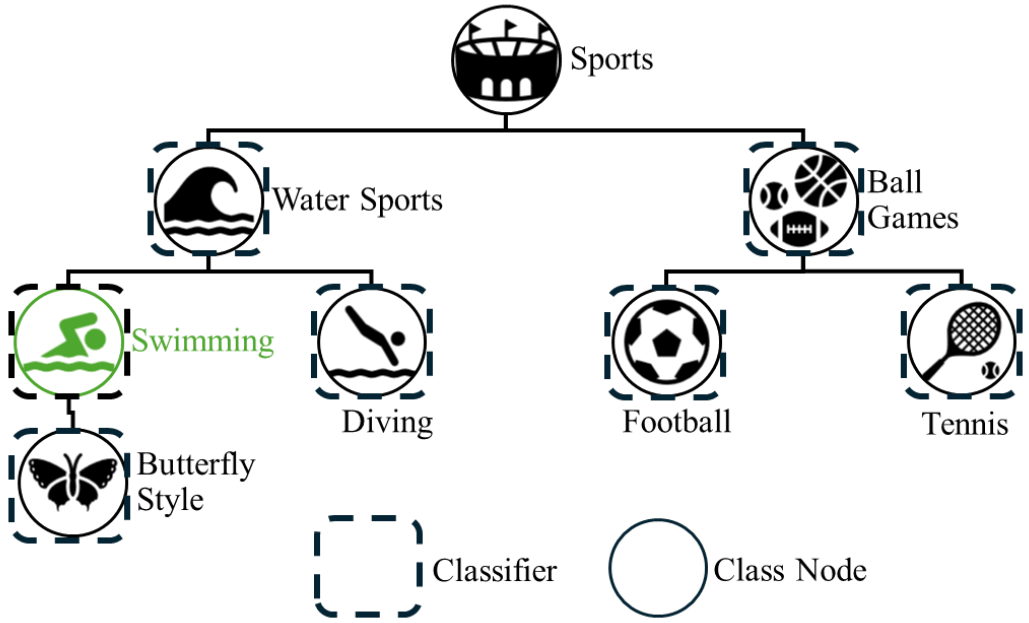
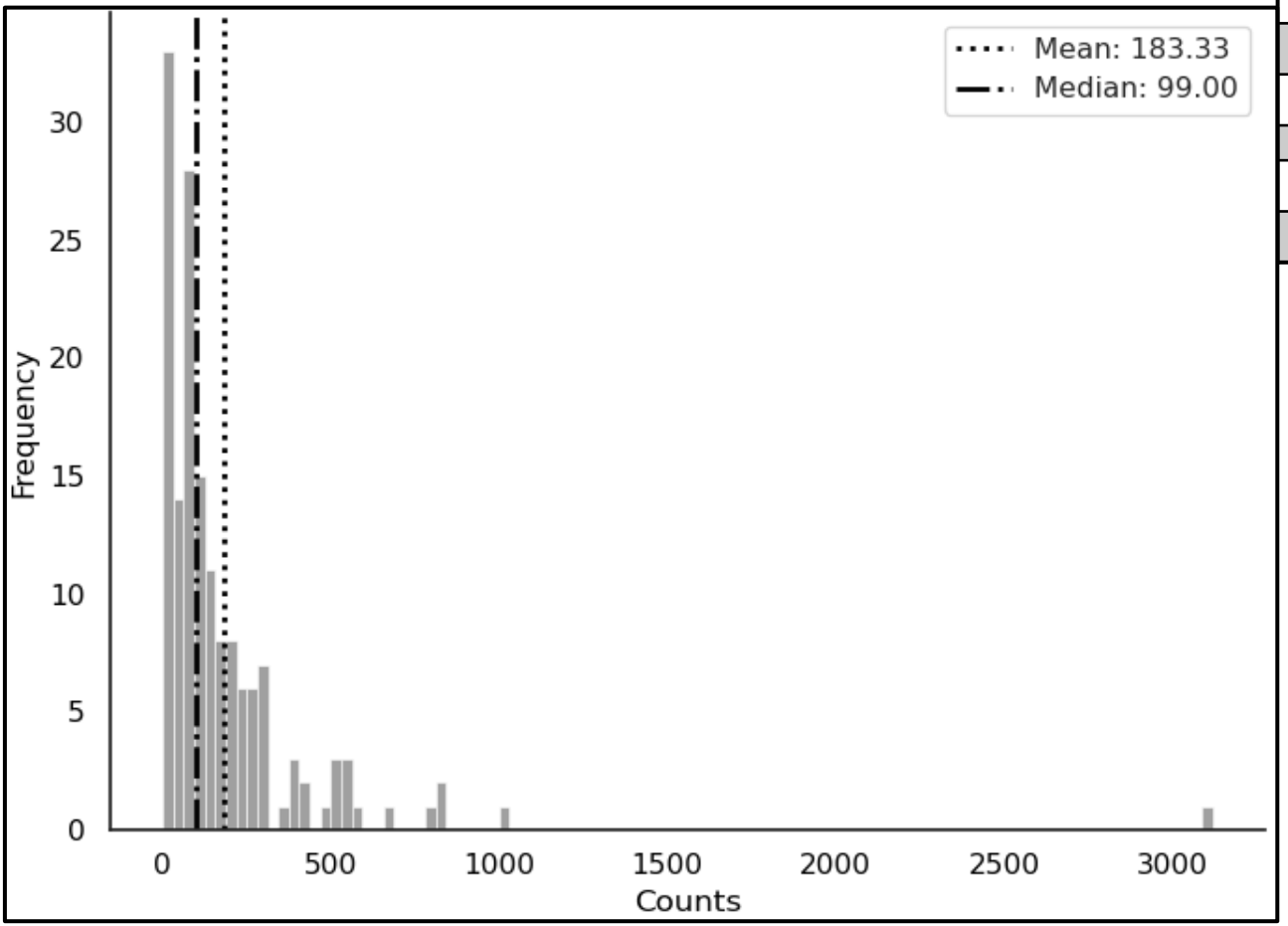


Illustration of Local Classifier Per Node Approach

Data: firm-year (2014.01-2023.12) accounting data from companies listed in the United States

Source: Compustat in WRDS (https://wrds-www.wharton.upenn.edu/)



Distribution of Numbers of Value Counts per Label Path

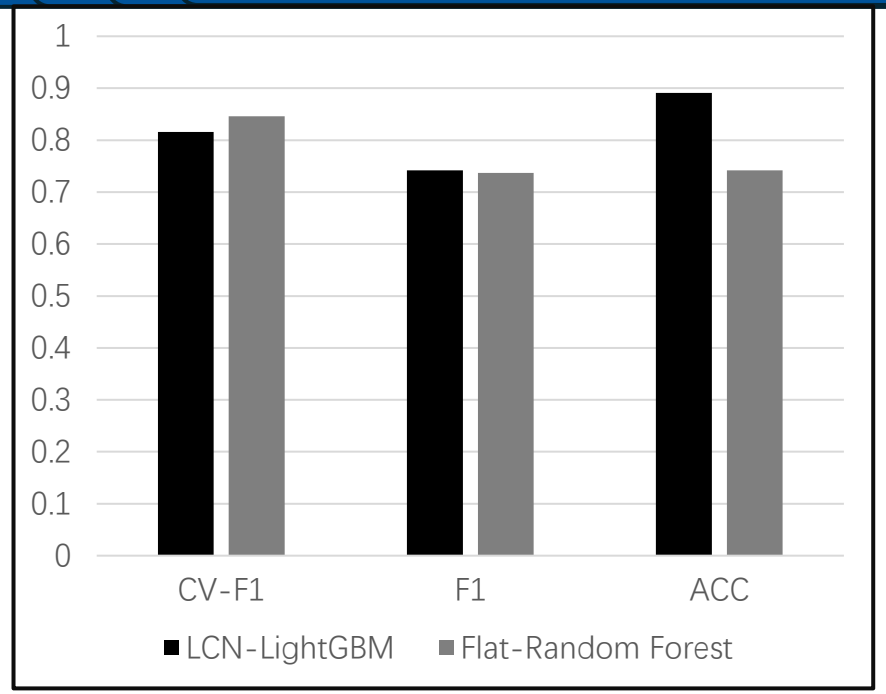
Criteria	Number of observations discarded	Number of observations remaining
Total Number available from Compustat	/	128, 646
Missing or invalid label	36, 261	92, 385
Duplicated observation	10, 537	81, 848
Total assets missing or <0	2, 463	79, 385
Balance sheet items >= total assets	10, 984	68, 401
Balance sheet items < 0	795	67, 606
Sales <= 0	7, 348	60, 258
Inventory <= 0	19, 402	40, 856
Final number of observations	/	40, 856

Data Screening Process

Data imbalance

- The most popular label path has over 3000 observations, while the least popular label path only has one.
- Half of the label paths have fewer than 100 observations.

Solution: full depth hierarchical classification oversampling (HROS-FD) techniques developed by Pereira et al. (2021) .



Model Performance

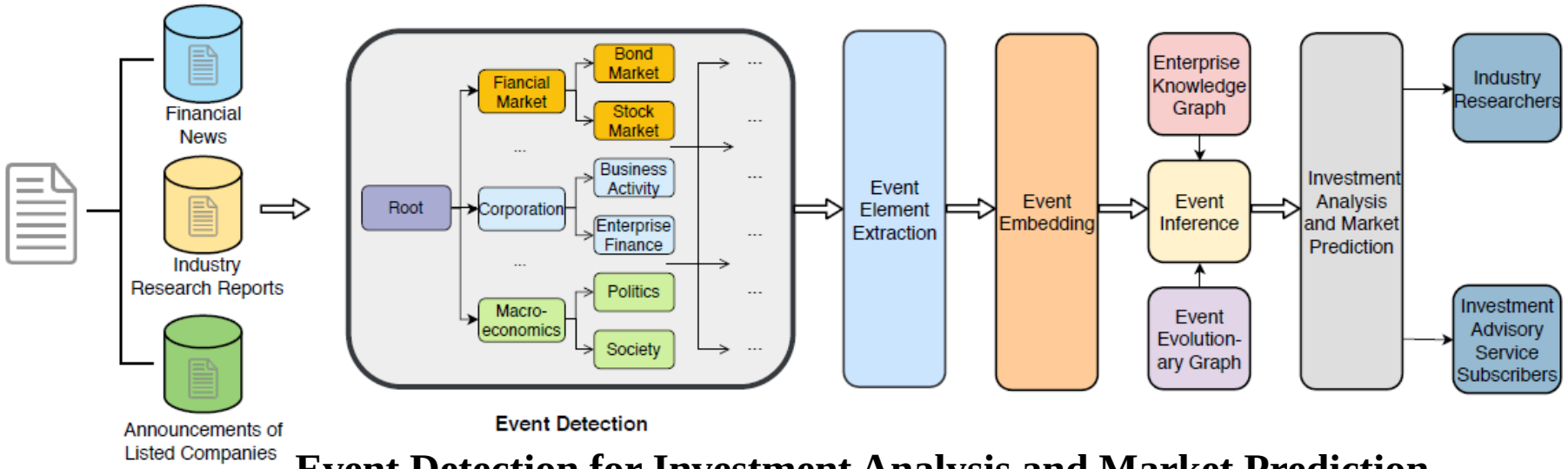
LCN-LightGBM has achieved almost 90% accuracy!

What have we achieved?

- Supervised learning + hierarchical structure + high accuracy
- Can be applied to predict other industry classification schemes (e.g., NAICS)
- Prove the usefulness of hierarchical classification in financial setting

Further application:

Below is another example of using hierarchical classification in financial setting. (This graph is credited to Liang et al. (2020))



Event Detection for Investment Analysis and Market Prediction