# Predicting Type of Work-related Fatal Accident Based on Knowledge Graph and Machine-Learning Methods

**Feiyu E[1, 4], Yiping Zeng[1], Fanyu Meng[1, 2], Pengjie Liu[3]& Lili Yang[1]**
**1** Department of Statistics and Data Science/ Southern University of Science and Technology/ Shenzhen, China,
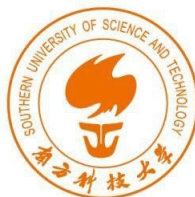**2** Academy for Advanced Interdisciplinary Studies/ Southern University of Science and Technology/ Shenzhen, China,
**3** Department of Computer Science & Engineering/ Southern University of Science and Technology/ Shenzhen, China,
**4** Faculty of Business and Economics/ The University of Hong Kong/ Hong Kong.

# Introduction

China's work-related injury insurance is controlled by the government, and accident cases are not available. Therefore, it is difficult to obtain data in the study of work-related accidents. As a result, the public's understanding of workplace injuries is not based on real data.

This study is based on the death accident case information related to work-related injury from 2011 to 2017 provided by Shenzhen Urban Public Safety and Technology Institute (SUPSTI).

✓ Descriptive statistics of data.
✓ Intuitive visualization of knowledge graph based on Neo4j and the application of the search function.
✓ Prediction of type of accident of work-related injury events based on machine learning algorithm.

# Introduction

**Objectives :**

1. Descriptive statistical analysis.

2. Knowledge graph.

3. Machine-learning-based prediction model.

# Experimental Design

1. Through the statistical analysis, the dataset adopted in this study can truly reflect the identification situation of insurance claims for work-related injury (death) in Shenzhen.

2. Through the knowledge graph tool, the structured data is converted into an entity-relation diagram form. It enables us more intuitively understand the correlation between the entities of industrial injury insurance. At the same time, it can realize the relationship search function.

3. After applying machine learning classification models, the type of accident can be predicted based on historical data. Using the company type, location, type of work, and other information to be the indicators of prediction, the trained model is used to do the prediction. It helps governments and enterprises to have a more targeted prevention of work-related injury.
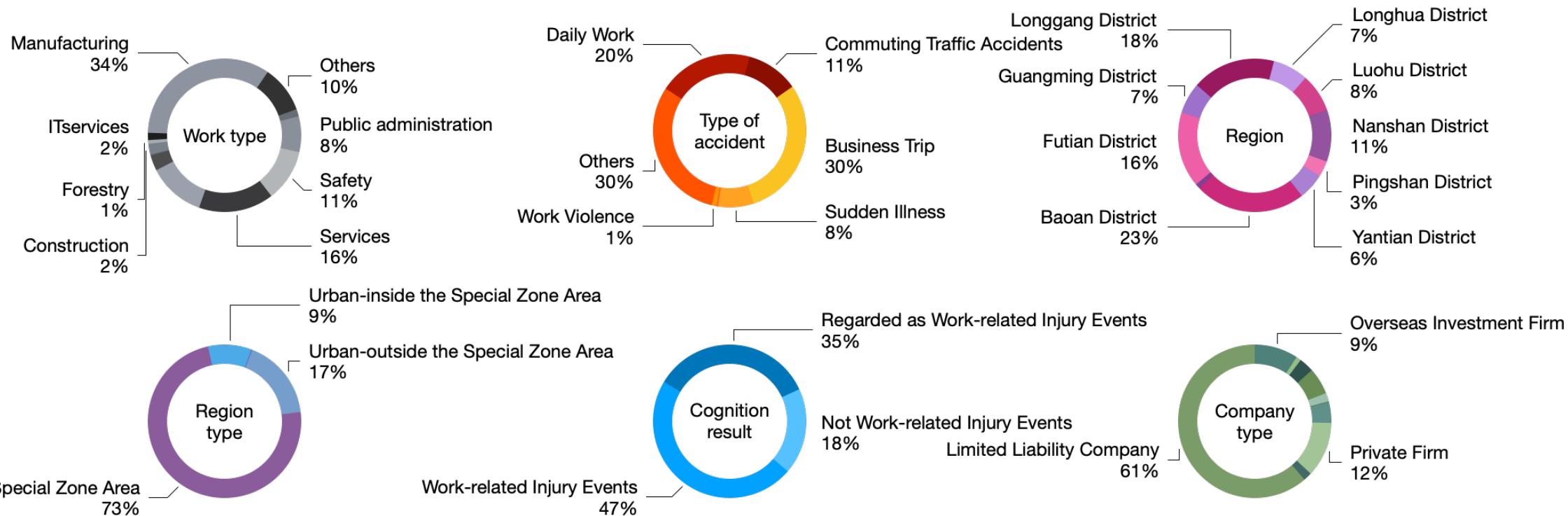
# Data

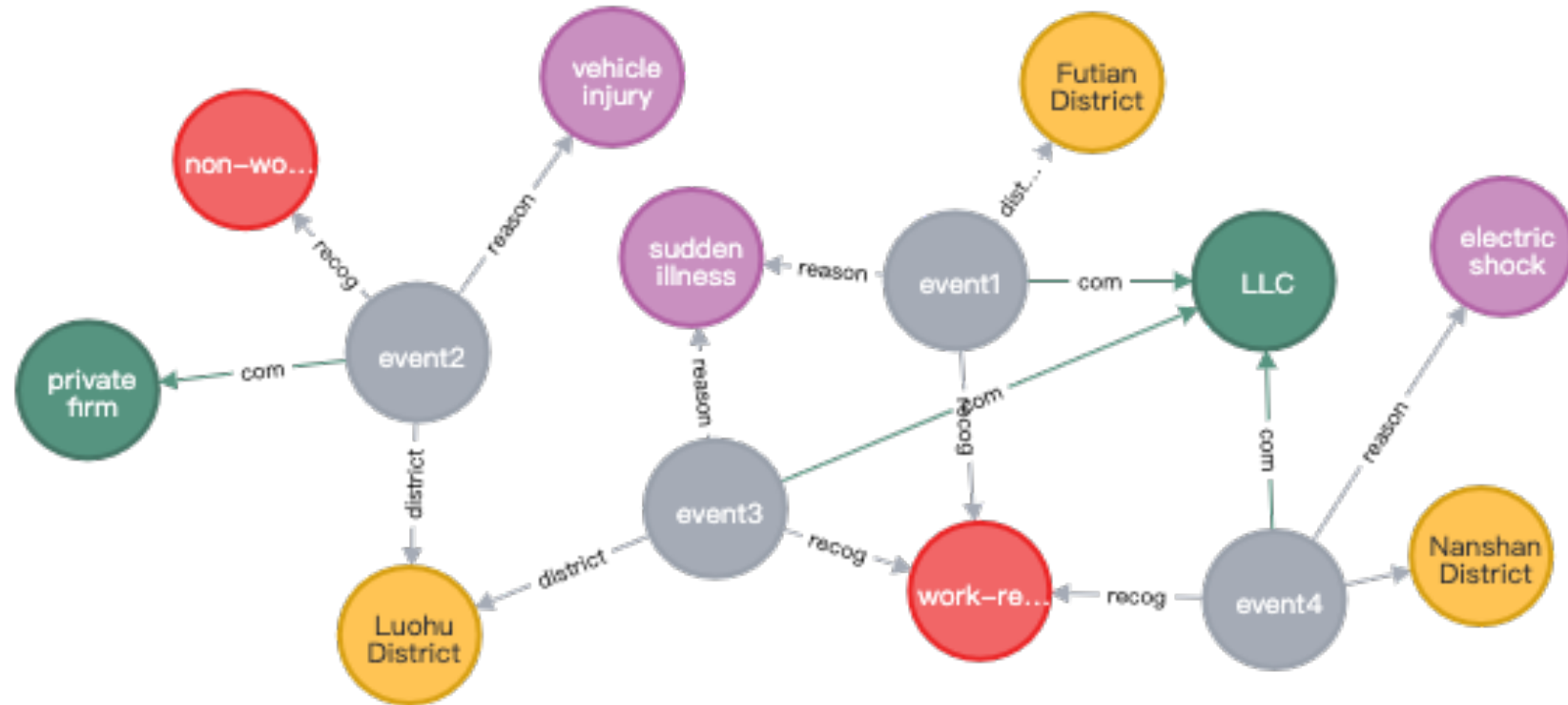| comtype | regiontype | accitype- | accitype | hurtreason | street | region | COGNRESU | persex | worktype | year | month | weekday |
|---------|-----------|-----------|----------|------------|--------|--------|----------|--------|----------|------|-------|---------|
| 有限责任公司 | 特区外农村 | 职业病 | 职业病 | 职业病 | 桃源 | 南山区 | 1 | 女 | Services | 2010 | 8 | 1 |
| 私有企业 | 特区外农村 | 道路车辆伤害 | 车辆伤害 | 其它原因 | 清水河 | 罗湖区 | 1 | 男 | Manufacturing | 2010 | 10 | 1 |
| 有限责任公司 | 特区外农村 | 道路车辆伤害 | 车辆伤害 | 上下班交通事 | 清水河 | 罗湖区 | 1 | 男 | Services | 2010 | 11 | 1 |
| 国有企业 | 特区内城镇 | 突发疾病 | 突发疾病 | 突发疾病 | 梅林 | 福田区 | 1 | 男 | Public administrati | 2010 | 11 | 1 |

The dataset acquired from SUPSTI, China contains 3470 death cases with regard to work- related injury insurance claimed in Shenzhen in 2011-2017.

Features are used including the type of company, the gender of the casualty, the type of region, the cause of injury, the street, and time (year, month and weekdays or not) of the accident. The work types are classified into 12 categories according to the types of work provided by Ehsani (2013). The causes of injury are clustered into four categories according to the relevant statutory circumstances.
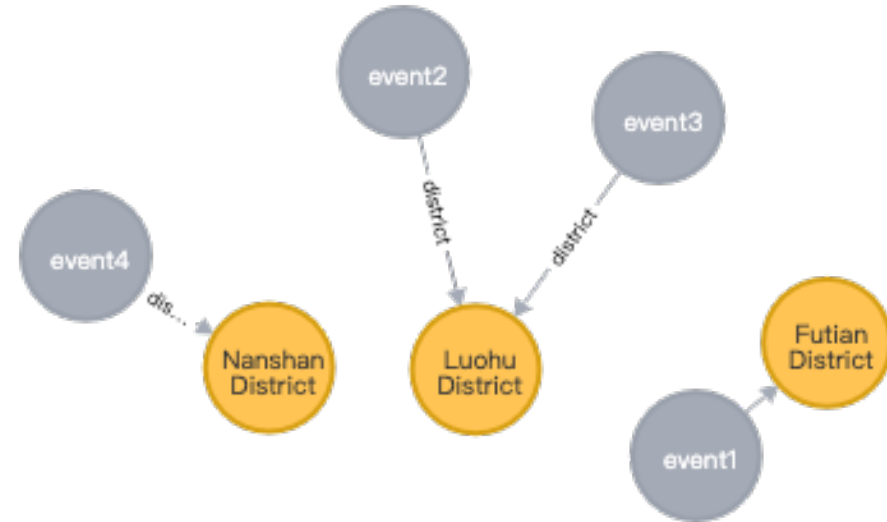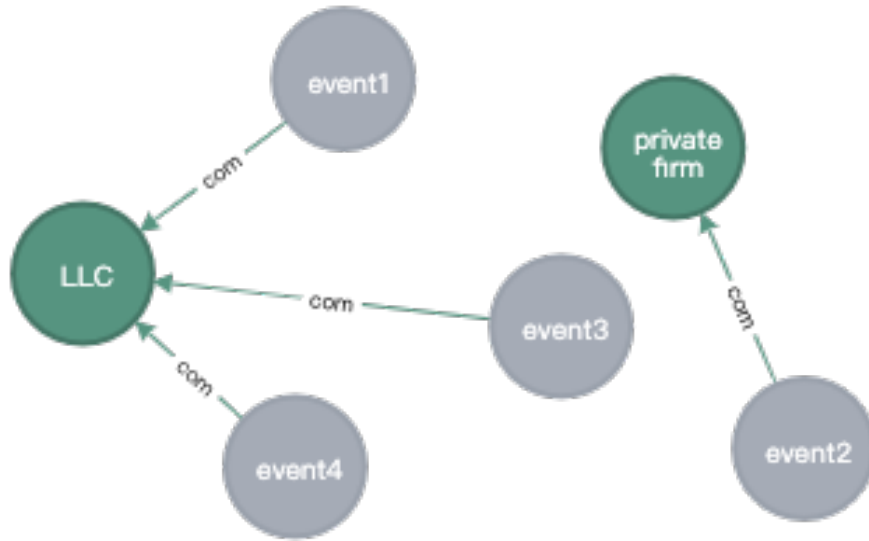
# Statistics



**Work type**
- Manufacturing 34%
- Others 10%
- Public administration 8%
- Safety 11%
- Services 16%
- Construction 2%
- Forestry 1%
- ITservices 2%

**Type of accident**
- Daily Work 20%
- Commuting Traffic Accidents 11%
- Business Trip 30%
- Sudden Illness 8%
- Work Violence 1%
- Others 30%

**Region**
- Longgang District 18%
- Longhua District 7%
- Luohu District 8%
- Nanshan District 11%
- Pingshan District 3%
- Yantian District 6%
- Baoan District 23%
- Futian District 16%
- Guangming District 7%

**Region type**
- Urban-inside the Special Zone Area 9%
- Urban-outside the Special Zone Area 17%
- Rural-outside the Special Zone Area 73%

**Cognition result**
- Regarded as Work-related Injury Events 35%
- Not Work-related Injury Events 18%
- Work-related Injury Events 47%

**Company type**
- Overseas Investment Firm 9%
- Private Firm 12%
- Limited Liability Company 61%
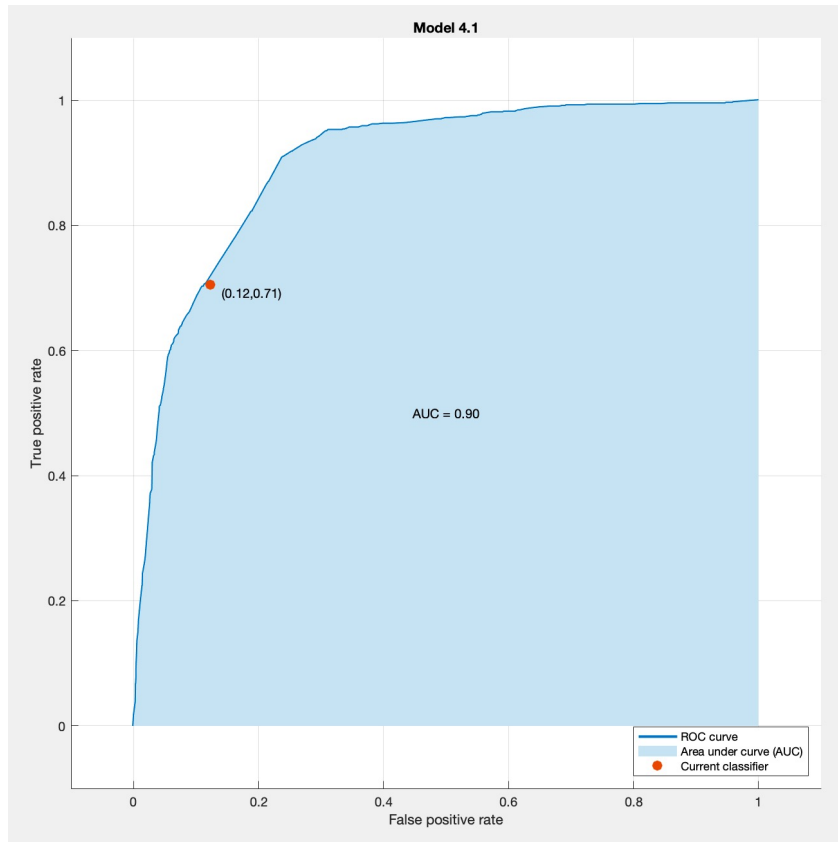
# Knowledge Graph

# Knowledge Graph

# Results

Machine learning classification algorithms including Decision Tree, SVM, KNN, and the ensemble algorithm AdaBoost (Cao et al., 2013) are selected and compared.

| Model | | Prediction speed | Training time |
|---|---|---|---|
| ☆ Tree<br>change: Medium Tree | Accuracy: **67.3%**<br>10/10 features | 200000 obs/sec | 0.8189 sec |
| ☆ SVM<br>change: Medium Gaussian SVM | Accuracy: **66.9%**<br>10/10 features | 19000 obs/sec | 6.7146 sec |
| ☆ KNN<br>change: Coarse KNN | Accuracy: **64.6%**<br>10/10 features | 8700 obs/sec | 1.6502 sec |
| ☆ Ensemble<br>change: Boosted Trees | Accuracy: **67.6%**<br>10/10 features | 26000 obs/sec | 4.5176 sec |

# Results

ROC ：

Confusion Matrix ：

# Conclusion

✓ This research illustrates the statistics of the fatal work-related injury cases in Shenzhen, China and displays them with features using Knowledge Graph with a search function of relevant entities.

✓ AdaBoost is the superior model among selected machine learning classification methods to predict the most relevant work-related accidents to a certain type of company under a given circumstance.

✓ Advice could be further made upon utilizing the prediction model for enterprises to make precautional plans to minimize the risk of work-related fatal accidents and injuries.

# Thanks

- **Questions (u3005658@connect.hku.hk)**

Founded by

- **The National Natural Science Foundation of China**
  Project No. 71771113.
- **The National Key R&D Program of China**
  Project No. 2018YFC0807000, and No. 2019YFC0810705