InCIT2020 October 21-22 @BUU, ChonBuri
The 5th International Conference on Information Technology: InCIT2020

Non-Communicable Disease Classification using Multi-Label Techniques

Heart of the Northeast
www.msu.ac.th

Worawith Sangkatip, Jiratta Phuboon-ob

Research Center of Information Technology for the future,

Department of Information technology, Faculty of Informatics,

Mahasarakham University, Thailand

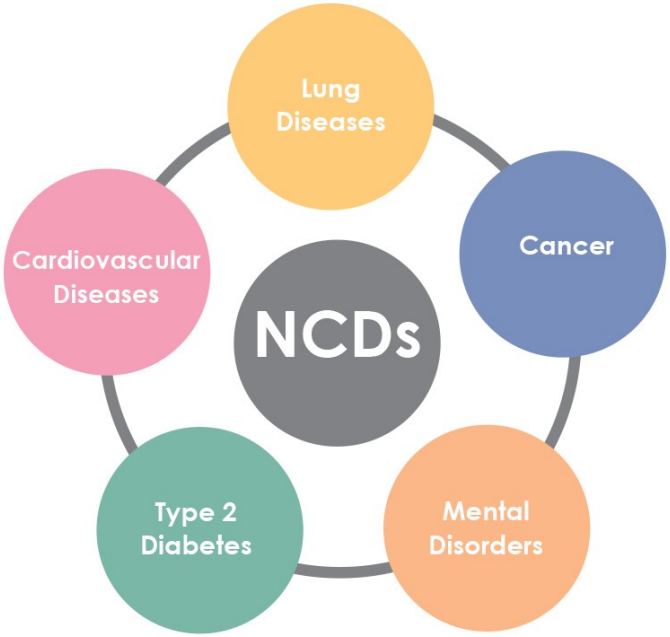
Outline

- Introduction
- Research Problem
- Research Contributions
- Related Work
- Methodology
- Dataset / Data Preprocessing
- Evaluation Measures
- Experimental Setup / Experimental Result
- Conclusion / Feature Work

Introduction

- Non-communicable diseases : NCDs
- NCDs are not transmissible directly from one person to another.
- World Health organization (WHO) reports that NCDs cause 41 million deaths each year, accounting to 71% of deaths in the world.
- In Thailand, NCDs cause several deaths, especially for persons who are over 30 years old.

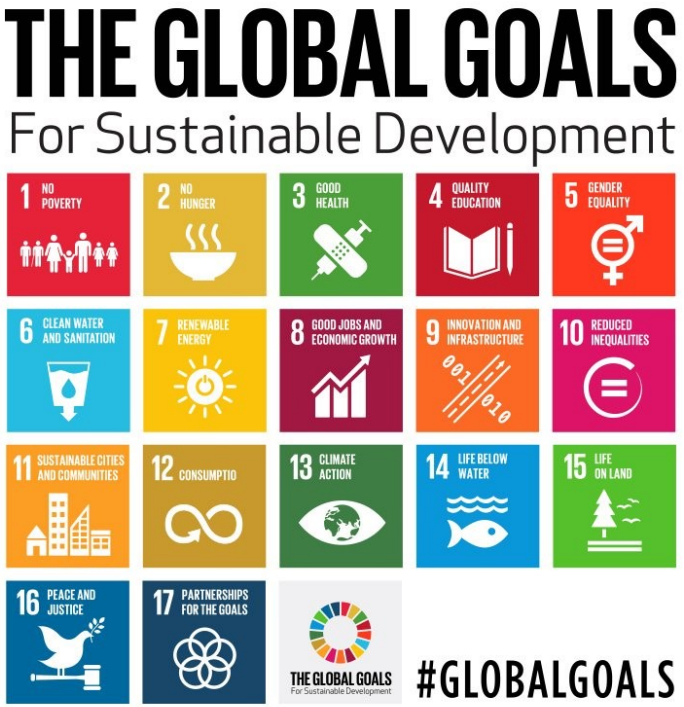
Research Problems



- NCDs patients always have multi-morbidities. For example, diabetic patients usually have hypertension symptom.
- Classification NCDs disease on patients who diagnosed with multi morbidity illness.

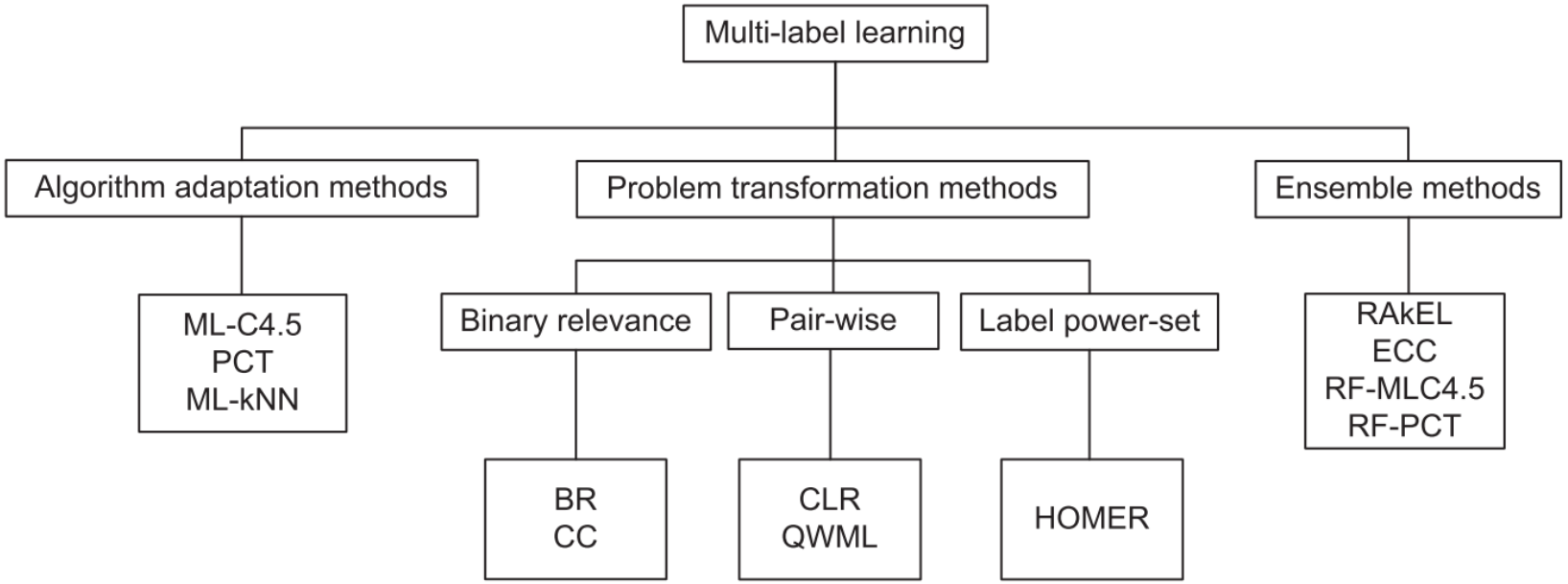
Research Contributions

- Aims to classify NCDs disease on patients who diagnosed with multi morbidity illness.
- Challenging issues in multi-label classification research.
- Results provided predictive algorithms with the highest level of accuracy for the multiple NCDs patients.



Related Work

- Multi-label learning has been presented by Boutell et al.(2004), Tsoumakas and Katakis (2007), Madjarov et al.(2012) Three categories of method.



Ref : G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, “An extensive experimental comparison of methods for multi-label learning,” in *Pattern Recognition*, 2012, vol. 45, no. 9, pp. 3084–3104

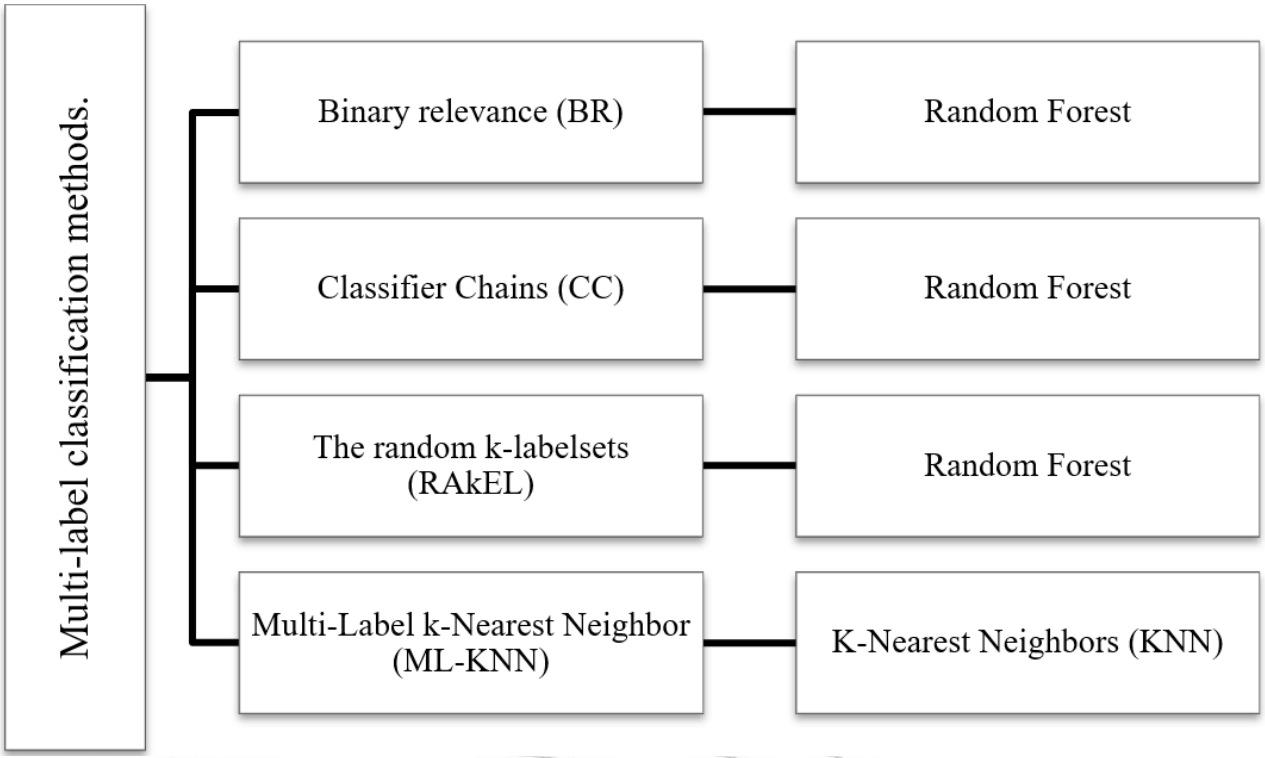
Related Work (Cont.)

- Related Work for Disease Diagnosis Domain

Author	Propose	Dataset
(Li et al. 2016b)	A Multi-Label Problem Transformation Joint Classification (MLPTJC)	physical examination records 110,300. 3 disease data, including diabetes, hypertension, and fatty liver,
(Li et al. 2017a)	a novel Ensemble Label Power-set Pruned datasets Joint Decomposition (ELPPJD)	physical examination records 110,300. 6 normal chronic diseases. They are hypertension, diabetes, fatty liver, cholecystitis, heart disease, and obesity
(Zhang et al. 2019)	A Novel Deep Neural Network Model for Multi-Label Chronic Disease Prediction	physical examination records 110,300. from about 80,000 anonymous patients, 3 chronic diseases. records: hypertension (H), diabetes (D), and fatty liver (FL).

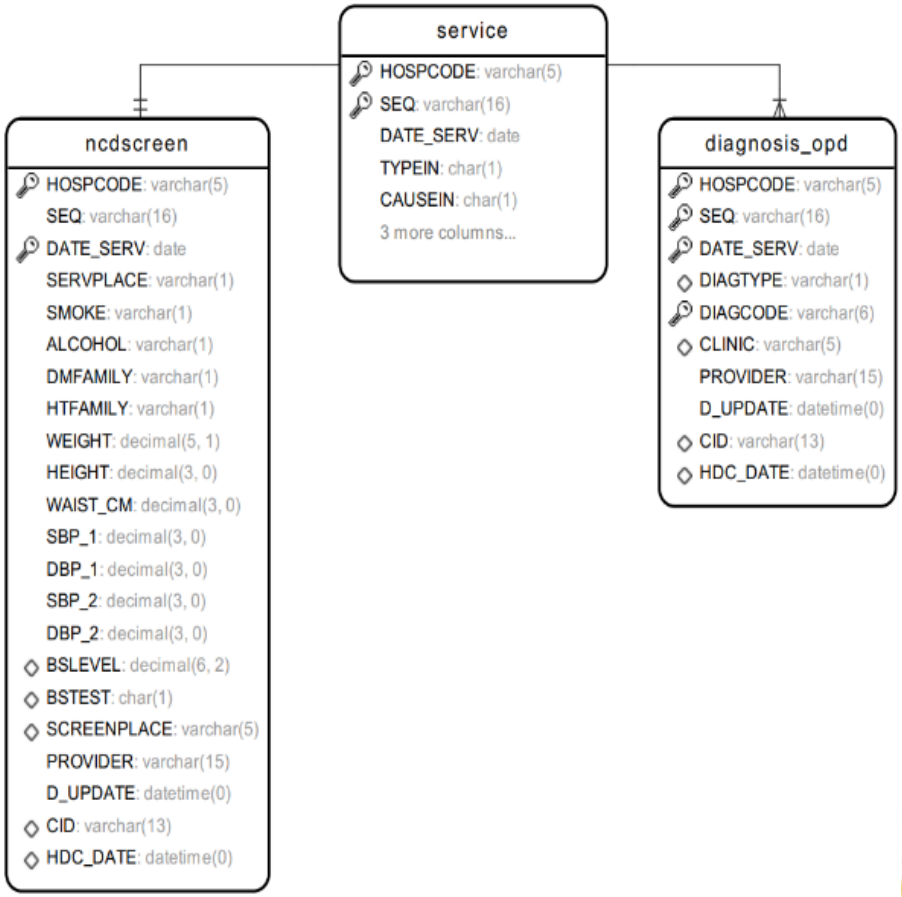
Methodology

- 4 methods, i.e. Binary relevance (BR) [4], Classifier Chains (CC) [8], The random k-labelsets (RAkE) [9] and Multi-Label k-Nearest Neighbor (ML-KNN) [10]



Datasets

- Dataset collected Electronic health record from Suddhavej Hospital.
- Data for this experiment was 19,554 medical examination collected during 2014 until 2019
- The focus 4 diseases : Diabetes, Hypertension, Cardiovascular and Stroke



Data Preprocessing

- The datasets are composed with two tables, according to NCDSCREEN and Diagnosis_OPD
- NCDSCREEN records the information from the patients.
- Diagnosis_OPD is a table collected from the diagnostic results of patients who use the medical services

Data Preprocessing (Cont.)

- Filter : Diagnosis code (ICD10)

Disease	Diagnosis code (ICD10)
diabetes	E10, E11, E12, E14
hypertension	I10, I11, I12, I13, I14, I15
cardiovascular	I20, I21, I22, I23, I24, I25
stroke	I60, I61, I62, I63, I64

- Data Integration

PID	x_1	x_2	x_m	diabetes	hypertension	cardiovascular	stroke
P_1	x_{11}	x_{12}	x_{1m}	1	1	0	0
P_2	x_{21}	x_{22}	x_{2m}	0	1	1	0
P_3	x_{31}	x_{32}	x_{3m}	0	1	0	1
P_n	x_{n1}	X_{n2}	X_{nm}	1	1	0	1

Data Preprocessing (Cont.)

- The attributes are classified into thirteen groups and data types are demonstrated in table.

Attributes	Description	Data Type
SMOKE	Smoking history	Nominal
ALCOHOL	Alcoholic drinking history	Nominal
DMFAMILY	Diabetes history in direct relatives	Nominal
HTFAMILY	Hypertension history in direct relatives	Nominal
WEIGHT	Weight	Numeric
HEIGHT	Height	Numeric
WAIST_CM	Waist circumference	Numeric
SBP_1	Systolic Blood Pressure: SBP 1st test	Numeric
DBP_1	Diastolic Blood Pressure: DBP 1st test	Numeric
SBP_2	Systolic Blood Pressure: SBP 2nd test	Numeric
DBP_2	Diastolic Blood Pressure: DBP 2nd test	Numeric
BSLEVEL	Blood sugar levels	Numeric
BSTEST	Methods of checking blood sugar	Nominal
Label_Diabetes	Diabetes diagnosis (0=negative, 1=positive)	Nominal
Label_Hypertension	Hypertension diagnosis (0= negative, 1= positive)	Nominal
Label_Cardiovascular	Cardiovascular diagnosis (0= negative, 1= positive)	Nominal
Label_Stroke	Stroke diagnosis (0= negative, 1= positive)	Nominal

Data Preprocessing (Cont.)

- The final summary of the dataset used in the experiment.
  - Label Cardinality (Card)

$$Label-cardinality=\frac{1}{N}\sum_{i=1}^{N}|Y_i|$$

  - Label density (Dens)

$$Label-density=\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i|}{L}$$

Instances	Features	Label	Label set	Card	Dens
19,554	13	4	15	0.151	0.038

Evaluation Measures

- Accuracy

$$\text{Avg accuracy} = \frac{\sum_{i=1}^{l}(\text{TP}_i + \text{TN}_i)/(\text{TP}_i + \text{FP}_i + \text{TN}_i + \text{FN}_i)}{l}$$

- Hamming Loss

$$Hamming\ loss = \frac{1}{|N| \cdot |L|} \sum_{i=1}^{|N|} \sum_{j=0}^{|L|} xor(y_{ij}, z_{ij})$$

Experimental Setup

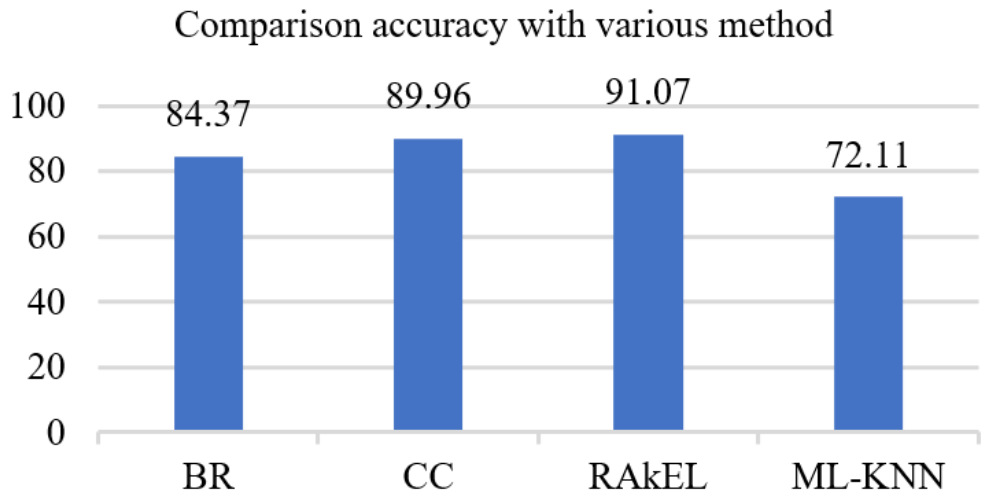
- Multi Label Method :
  - Binary relevance (BR)
  - Classifier Chains (CC)
  - The random k-labelsets (RAkEL)
    - subset (k) is 3, the number of subsets (m) is 10.
  - Multi-Label k-Nearest Neighbor (ML-KNN)
    - number of neighbors (k) is 3.
- Parameters in Random Forest : num trees is 1000, max depth is 0 (unlimited depth), num features is 0.

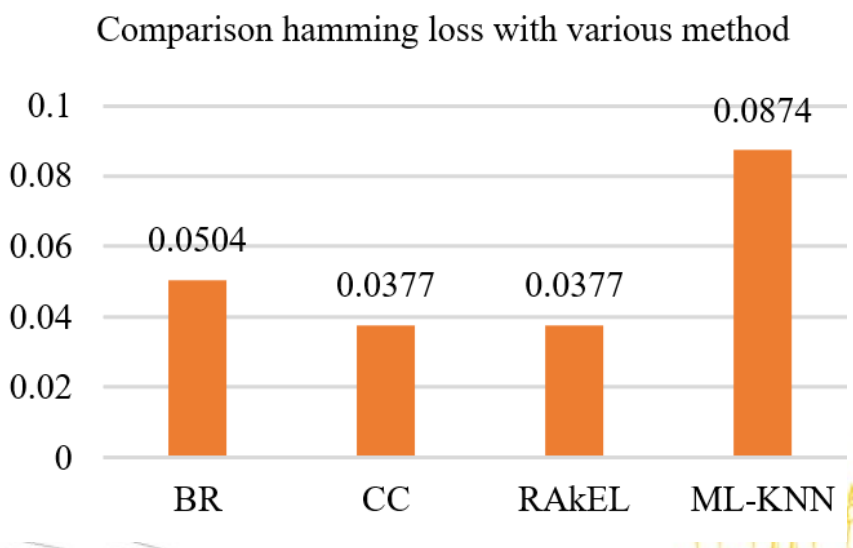
Experimental Setup

- The Tool is Meka Software, which is an extension of the Weka program, were used in BR, CC, RAkEL methods.
- ML-KNN method used MULAN framework.
- Evaluation model with 10-fold cross validation.
- Evaluation Measures : Accuracy, Hamming loss.

Experimental Result

Methods	Accuracy (%) ± S.D.	Hamming Loss ± S.D.
BR	84.37 ± 0.0065	0.0504 ± 0.0024
CC	89.96 ± 0.0061	**0.0377** ± 0.0022
RAkEL	**91.07** ± 0.0074	**0.0377** ± 0.0025
ML-KNN	72.11 ± 0.0084	0.0874 ± 0.0030





Conclusion / Future Work

- The RAkEL method is the most effective method with an accuracy rate of 91.07%, the highest rate compared with the other three methods.
- Future work is Develop multi-label classification method using deep learning algorithms for NCDs data. Which is anticipate achieving higher accuracy.

Acknowledgement

- This research is supported Suddhavej Hospital, Faculty of Medicine, Mahasarakham University, who provides the NCDs' screening data.

Q & A