THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

HEALTHCARE DATA MINING FROM MULTI-SOURCE DATA

Ling Chen

Master of Computer Science

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2017

School of Information Technology and Electrical Engineering

Abstract

The "big data" challenge is changing the way we acquire, store, analyse, and draw conclusions from data. How we effectively and efficiently "mine" the data from possibly multiple sources and extract useful information is a critical question. Increasing research attention has been drawn to healthcare data mining, with an ultimate goal to improve the quality of care. The human body is complex and so too the data collected in treating it. Data noise that is often introduced via the collection process makes building Data Mining models a challenging task.

This thesis focuses on the classification tasks of mining healthcare data, with the goal of improving the effectiveness of health risk prediction. In particular, we developed algorithms to address issues identified from real healthcare data, such as feature extraction, heterogeneity, label uncertainty, and large unlabeled data.

The three main contributions of this research are as follows. First, we developed a new health index called Personal Health Index (PHI) that scores a person's health status based on the examination records of a given population. Second, we identified the key characteristics of the real datasets and issues that were associated with the data. Third, we developed classification algorithms to cope with those issues, particularly, the label uncertainty and large unlabeled data issues.

This research takes one step forward towards scoring personal health based on mining increasingly large health records. Particularly, it pioneers exploring the mining of GHE data and tackles the associated challenges. It is our anticipation that in the near future, more robust data-mining-based health scoring systems will be available for healthcare professionals to understand people's health status and thus improve the quality of care.

Declaration by Author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Journal paper:

- Ling Chen, Xue Li, Quan Z. Sheng, Wen-Chih Peng, John Bennett, Hsiao-Yun Hu, and Nicole Huang, " Mining Health Examination Records − A Graph-based Approach", *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 28, iss. 9, pp. 2423−2437, 2016.

- Ling Chen, Xue Li, Yi Yang, Hanna Kurniawati, Quan Z. Sheng, Hsiao-Yun Hu, and Nicole Huang, "Personal health indexing based on medical examinations: A data mining approach," *Decision Support Systems (DSS)*, vol. 81, pp. 54−65, 2016.

Conference paper:

- Ling Chen, Xue Li, Sen Wang, Hsiao-Yun Hu, Nicole Huang, Quan Z. Sheng, and Mohamed A. Sharaf, "Mining Personal Health Index from Annual Geriatric Medical Examinations," in *IEEE International Conference on Data Mining (ICDM)*, 2014, pp. 761−766.

Poster:

- Ling Chen, Xue Li, Quan Z. Sheng, and Mohamed A. Sharaf, "Mining Personal Health Index from Annual Geriatric Medical Examinations," in *The 2014 Australasia Database Conference PhD School on Big Data*, The University of Queensland, Australia, 2014. [**Best Research Poster Awards - Joint First Prize**]

Publications included in this thesis

Ling Chen, Xue Li, Sen Wang, Hsiao-Yun Hu, Nicole Huang, Quan Z. Sheng, and Mohamed A. Sharaf, "Mining Personal Health Index from Annual Geriatric Medical Examinations," in IEEE International Conference on Data Mining (ICDM), 2014, pp. 761−766. - incorporated as Chapter 3.

Contributor	Statement of contribution
Ling Chen (Candidate)	Data preprocessing (100%) Algorithm and experiment design (78%) Paper writing (75%)
Xue Li	Algorithm and experiment design (20%) Revising the paper (20%)
Sen Wang	Algorithm and experiment design (2%)
Hsiao-Yun Hu	Coordination of the data
Nicole Huang	Professional advice from public health
Quan Z. Sheng	Revising the paper (5%)
Mohamed A. Sharaf	General idea discussion

Ling Chen, Xue Li, Yi Yang, Hanna Kurniawati, Quan Z. Sheng, Hsiao-Yun Hu, and Nicole Huang, "Personal health indexing based on medical examinations: A data mining approach," *Decision Support Systems*, vol. 81, pp. 54−65, 2016. - incorporated as Chapter 4.

Contributor	Statement of contribution
Ling Chen (Candidate)	Data preprocessing (100%) Algorithm and experiment design (70%) Paper writing (75%)
Xue Li	Algorithm and experiment design (10%) Revising the paper (15%)
Yi Yang	Algorithm and experiment design (20%)
Hanna Kurniawati	Revising the paper (5%)
Quan Z. Sheng	Revising the paper (5%)
Hsiao-Yun Hu	Coordination of the data
Nicole Huang	Professional advice from public health

Ling Chen, Xue Li, Quan Z. Sheng, Wen-Chih Peng, John Bennett, Hsiao-Yun Hu, and Nicole Huang, " Mining Medical Examination Records − A Graph-based Approach", *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 28, iss. 9, pp. 2423−2437, 2016. - incorporated as Chapter 5.

Contributor	Statement of contribution
Ling Chen (Candidate)	Data preprocessing and generation (100%) Algorithm and experiment design (85%) Paper writing (75%)
Xue Li	Algorithm and experiment design (15%) Revising the paper (15%)
Quan Z. Sheng	Revising the paper (5%)
Wen-Chih Peng	Revising the paper (5%)
John Bennett	Clinical advises on experiment design Revising the paper
Hsiao-Yun Hu	Coordination of the data
Nicole Huang	Professional advice from public health

Contributions by others to the thesis

For all the published research works included in this thesis, my principle advisor Professor Xue Li provided very helpful insight. Especially, he assisted in providing guidance for problem formulation, the refinement of the ideas and the pre-submission edition, as well as reviewing and polishing the presentation. My co-advisor, Dr Mohamed A. Sharaf, assisted in refinement of research problems and reviewing the presentation. Associate Professor Yi Yang also provided insight in machine learning algorithms, especially regarding the problem of learning with uncertain labels. Dr. Sen Wang assisted in providing technical suggestions for experiments.

Professor Michael Q. Z. Sheng often gave insightful comments while revising the papers. The comments and suggestions from General Practitioner Dr. John Bennett at UQ Health Service have shaped our problem formulations and advanced our understanding of the experimental results. Professor Nicole Huang and Dr. Hsiao-Yun Hu have supported in our understanding of the real datasets as well as practical concerns from the aspect of Public Health.

Statement of parts of the thesis submitted to qualify for the award of another degree

"None"

Acknowledgments

The journey towards the completion of a Ph.D degree has been a valuable learning experience, for which I am very grateful. I would not have achieved this milestone without the support of many people.

I would like to express my special gratitude to my principle advisor and mentor Professor Xue Li, for his guidance, support, and his sometimes surprisingly brilliant ideas that kept me excited about this research direction. I also thank my associate advisor Dr Mohamed Sharaf for his comments and suggestions. For all committee members, especially Associate Professor Udantha Abeyrantne as the chair of committee, I would like to thank you all for the suggestions you offered to improve my thesis.

I must express my gratitude for our collaborators in Taiwan, Professor Nicole Huang and Dr. Hsiao-Yun Hu. Without their support I would not have had access to the real health examination datasets which have been integral for this research.

To other members of DKE group especially my Room 78-636 officemates, Sayan Unankard, Hongyun Cai, and Xuefei Li, I am indebted to you for your help and encouragement. I would like to thank those who helped with proof-reading, especially Mr. Ross Hohl.

To my beloved parents, thank you very much for always supporting me and encouraging me in the ups and downs of this long journey.

The University of Queensland and the School of ITEE granted me the Australian Postgraduate Scholarship and the UQ Advantage Scholarship that allowed me to undertake this research, as well as other scholarships that allowed me to attend conferences which proved very beneficial.

The studies in this thesis were largely based on the de-identified data from the Taipei City Public Health Database at the Department of Health, Taipei City Government, and managed by Databank for Public Health Analysis (DoPHA). The interpretation and conclusions contained herein do not represent those of Department of Health, Taipei City Government, or DoPHA. The research is partially funded by the Australian Research Council Discovery Project ID DP140100104.

Keywords

personal health index mining, health examination records, classification with label uncertainty, classification with large unlabeled data, graph-based semi-supervised learning

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080109 Pattern Recognition and Data Mining, 90%

ANZSRC code: 111711 Health Information Systems, 10%

Fields of Research (FoR) Classification

FoR code: 0801, Artificial Intelligence and Image Processing, 60%

FoR code: 0806, Information Systems, 40%

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In early 2015, the United States government announced an initiative for precision medicine [30], which referred to prevention and treatment strategies that took individual variability into account. This was not a single event but signified an emerging trend of improving personalised healthcare by utilising increasingly large amounts of available health data as well as powerful computational tools for data analysis.

The field of research that explores Data Mining technology to meet the needs of the healthcare industry has a variety of names, such as Healthcare Data Mining, Data Mining in Healthcare [74], Medical Data Mining [110, 28], and Clinical Data Mining [62], depending on the emphasis of the types of data and applications. In general, digitally stored healthcare data, possibly collected from multiple information sources and in different forms, is called Electronic Health Record (EHR). An EHR covers a wide range of information about an individual, such as but not limited to, person demographics, observations, laboratory tests, diagnostic reports, treatments, therapies, prescriptions, and allergies [36]. By monitoring a person's EHRs over time, the longitudinal fluctuation of his or her health status can be tracked. Therefore the longitudinal EHRs of individuals provide a rich repository for healthcare data analysis and predictive modeling. However, developing effective models for healthcare applications is very challenging, especially due to noise, incompleteness, heterogeneity, and longitudinality of the real data.

In this chapter, we briefly introduce the motivation, background, challenges and goals, and our main contributions in healthcare data mining.

1.1 Motivation

Most developed countries have experienced dramatic growth in elderly populations from the beginning of this century. In recent years, this, together with the rising cost of healthcare has created an urgent need for improving predictions and efficient treatment. For example, the U.S. government spent $414.3 billion in elderly health care in 2011, $100 billion higher than the inflation-adjusted costs of 2001 [89]. Many developed countries, such as Australia [1], United Kingdom [9], and Taiwan [55], provide periodic General Health Examinations (GHEs) to their senior citizens as an integral part of preventive healthcare for the elderly. These programs enable continuous and comprehensive recording of a person's health status, as well as the tracking of his/her health changes. However, it is always a difficult task for healthcare professionals to provide an overall report on personal health after a comprehensive medical check-up has been performed because of hundreds of the parameters available to be considered.

One particular focus of preventive healthcare is risk assessment. The goal is to identify individuals at risk for further investigation or early treatment and intervention. Traditionally, risk assessments have been conducted manually by clinical professionals based on their expertise. These manual assessments have been constrained by the capacity of the human brain to process information within a limited time during the period of an appointment with a patient.

Many risk-scoring systems have been developed in the field of medicine to assist clinical decision-making, for example, the APACHE, SAPS, and MPM for Intensive Care Unit patients [42]. As a general practice in medical research, these methods have been defined based on factors selected with expert knowledge and validated via population-based studies [128].

With the advances in computing technology and the availability of EHRs, an increasing number of data mining and machine learning applications have been developed to support healthcare decision making [37, 134]. In recent years methods for clinical risk classification have been developed [122, 121, 127, 88, 92]. However, most existing studies have their focus on EHRs. GHE records and the unique challenges they pose have not yet been well explored. This gap has driven our research to advance risk prediction models for GHE records.

1.2 Background

To enable better understanding of our research and the challenges behind it, we introduce our datasets and discuss their characteristics below.

1.2.1 Datasets

The key datasets of this thesis contain a General Health Examination (GHE) dataset and a Cause of Death (COD) dataset. These datasets have been linked together via the common attribute Person ID, revealing the association between examination results and main Cause of Death. Records of participants with non-health-related COD were excluded in our preprocessing because the goal of our research is to build effective risk prediction models.

General Health Examination (GHE) Data

GHE is a de-identified dataset for elderly people with all private information, such as names, contact details, and birth dates removed. The dataset has 230 attributes, containing 262,424 check-ups of 102,258 participants aged 65 or above, collected during a period of six years, from 2005 to 2010. The overall ratio of male to female participants is 1.03:1. Each de-identified GHE record is represented by a Person ID and the examination results from a wide range of lab tests, physical examinations, the Brief Symptom Rating Scale (BSRS) mental health assessment, the Short Portable Mental Status Questionnaire (SPMSQ) cognitive function assessment, and (de-identified) demographics. Personal health-related habits are also available, such as exercise, eating, drinking, and smoking habits. Key attributes are listed in Table 1.1.

The data was collected during a standard annual health examination program for elderly people, run by the Taipei City Government, Taiwan (R.O.C.). Participants voluntarily took part in the program, and were encouraged to visit on a yearly basis. Data related to individual identification was removed before the dataset acquisition. The acquisition and processing of the data was approved by the Institutional Review Board (IRB) of the Taipei City Hospital, Taiwan (R.O.C.).

TABLE 1.1: Selected GHE attributes by categories

Type	Category	Attribute (example)
Patient Profile	Demographics	age, marital status, gender, education level, residential suburb
	Habits	reasons-for-taking-medicine, smoking, drinking, exercise, drink-milk, eat-vegetable, clean-teeth
Lab Tests	Biochemical	glu-ac, total cholesterol (tcho), thyroglobulin (tg), got, gpt, albumin (alb), thyroid stimulating hormone (tsh)
	Blood	red blood cell, white blood cell, plate, hematocrit (hct), mean corpuscular volume (mcv), mean corpuscular hemoglobin (mch), alpha-fetoprotein (afp), hemoglobin (hb)
	Urine	outlook, ph, protein, sugar, blood, red blood cell, white blood cell, pus cell, epithelium cell, casts
	Other	faecal occult blood test (fobt)
Examinations	Physical	weight, height, waist, systolic blood pressure, diastolic blood pressure, pulse rate
	External	neck, chest, heart, breast, abdomen, back, rectum, limbs, prostate
	Other	X-ray, EKG, cervical smear, abdominal ultrasound
Mental Health	BSRS	5 questions regarding nervousness, anger, depression, comparison with others, and sleep
Cognitive Function	SPMSQ	10 questions, e.g., current date, day of the week, where the person is situated, home address, age, year of birth, etc.

Cause of Death (COD) Data

The GHE dataset was linked to the Taiwan National Death Registry system using participants' identification numbers and then encrypted to provide de-identified secondary data maintained by the Department of Health of the Taipei City Government. We called this linked subset of data the Cause of Death (COD) dataset. The main causes of death were encoded with the WHO International Classification of Diseases [2], a standard medical ontology for disease classification. CODs were coded with the 9th ICD revision (ICD-9) from years 2005 to 2008 and the 10th ICD revision (ICD-10) from years 2009 to 2010. There were in total 522 ICD-9 codes and 925 ICD-10 codes used in the COD dataset. Attributes available from the linked information included a 3-4 digit ICD code for main cause of death and time of death in the format of month and year.

For the purpose of risk prediction, we only included those who passed away within three years of their last examination record. This left us with 7,569 participants with COD codes, which was about 7.4% of our dataset.

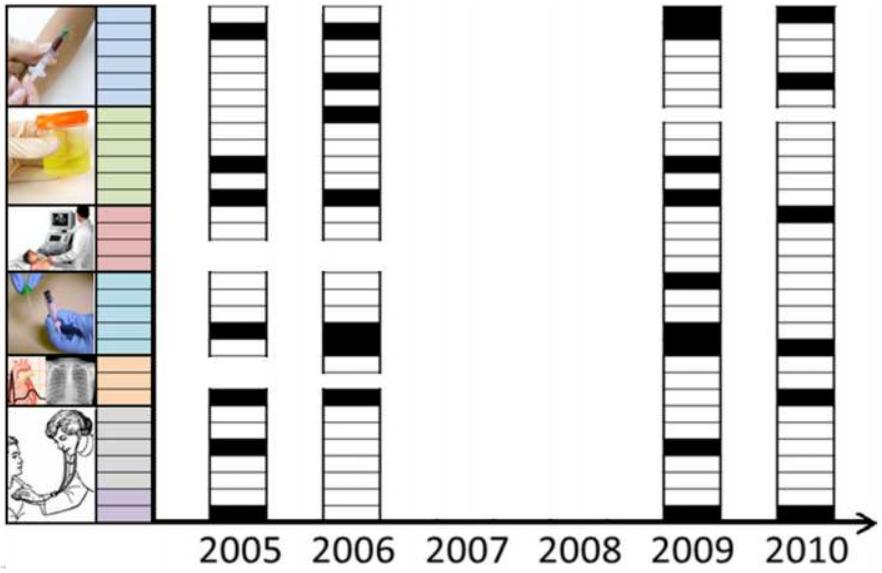

FIGURE 1.1: *An example of a person's health examination records.*

1.2.2 Data Characteristics

We have identified key characteristics of the above described GHE and COD datasets. These characteristics not only constitute the key challenges of our research, but also differentiate our work from existing studies on mining EHRs. Some characteristics have been associated with the GHE records; others were identified from the COD labels. In the following, we will discuss them in detail.

GHE Characteristics

To illustrate the key characteristics of GHE data, an example of the record sequence of a GHE participant is given in Figure 1.1. The horizontal axis is the time in years and the vertical axis indicates examination items of various categories. Cells in black mark the items whose results were abnormal at that year. The gaps between cells for a given year, say 2005, were the items not tested during that year; the years with no records, say 2007 and 2008, were the years that the person did not take the examination.

Three characteristics have been identified from our GHE dataset, as listed below.

- **Infrequency:** As GHE was offered on a yearly basis, the record sequences were infrequent, compared with time-series healthcare data, such as ECG and body movements collected from wearable sensor devices [88, 103, 109]. Time-series data from wearable devices were often

collected in the frequency of Hertz, so the problem was often how to extract more compact data representation to save computational cost [129]. By contrast, the infrequency of GHE data gave us relatively short sequences with various lengths.

- **Incompleteness:** Due to the voluntary nature of GHE, a person might have only taken a couple of non-consecutive GHEs in his or her lifetime. Therefore, the record sequences were incomplete along the timeline. Also, the health examination program may have changed over time so the examination items covered in the program would also have changed accordingly. As a result, GHE records were incomplete as well in the sense of missing observations of examination items in certain years.

- **Sparsity:** Clinical judgements were often based on abnormal findings such as observed symptoms, signs, or lab test results. From the perspective of abnormalities, GHE records, especially that of lab tests and mental health assessments, were likely to be sparse, because the majority of the results would have been likely to be normal.

- **Heterogeneity:** As shown in Table 1.1, GHE records were composed of results from a wide range of lab tests, physical examinations, cognitive and mental assessments, and personal habits (e.g., exercise and diet habits). They were often collected from different sources, using different types of machines and measurements, focusing on different parts of the human body, and reflecting different aspects of a person's health status. Therefore, this heterogeneity of GHE records should be respected and considered in the analysis and modeling of the data.

The above four characteristics of GHE data differentiate our work from two strains of traditional classification problems. On one hand, our problem was different from the traditional point-based classification, where an example is viewed as a point in a multi-dimensional space, represented as a feature vector. After all, in our case a person may have more than one record. On the other hand, our problem was also different from the traditional time-series classification problem, which focused on the handling of high frequency series, such as ECG series [129, 103, 109], because of the infrequency and incompleteness of our GHE records.

Label Characteristics

In addition to the characteristics of the GHE dataset, there are two characteristics associated with the COD labels that have been identified, namely label uncertainty and large unlabeled data, as illustrated below.

- **Label Uncertainty:** In a clinical context, a label can be a recorded illness, death, or a prescription of medicine. Label uncertainty has commonly been found in clinical judgments due to expert subjectivity and inadequate information [104]. Often, it is handled as noise, so the task has been to detect and correct such mislabeling [87, 107, 45]. However, in the case of multiple, non-exclusive medical conditions [82], such as comorbidity, it makes more sense to treat labels with degrees of certainty rather than forcing them to belong to one "true" class, because there is no such thing as a single true class in this kind of scenario.

  This issue applies to COD labels, because a participant might have comorbidity but only the main cause of death was recorded. Therefore, the recorded COD labels should be treated as with a degree of certainty. Note that if multiple labels are given at a point in time, it becomes a multi-label learning problem that predicts multiple labels at a time, which is a different problem from what we are dealing with here.

- **Large Unlabeled Data:** About 92.6% of the 102,258 participants in the GHE dataset did not have a COD label. The semantics of such "alive" cases varied from generally healthy to very ill, and anywhere in between. This implies that there is no ground truth available for the "healthy" cases, which meant we were dealing with large unlabeled data. In fact, the unlabeled data rate in our dataset was 92.6%.

These two characteristics associated with the COD labels constitute the key challenges of our research, which will be further elaborated in Section 1.3.

Differences from EHRs

Although GHE records, being also digitally stored health data, can be seen as a subset of EHRs, it is worth noting their differences. EHRs have been mainly collected for patient management and subsequent administrative use. The focus has been on the clinical care for a specific problem of a patient. Therefore the recorded data, at any given point in time, also had its focus on the specific problem. After all, given limited medical resources, a doctor would only order tests that are considered as necessary for further understanding an illness of a patient or confirming a diagnosis for the patient.

By contrast, GHE records often covered a rather comprehensive set of general health measures for regular surveillance and preventive purposes [86]. A large number measures were collected at a point in time during the examination program in a systematic way. This provides a comprehensive coverage of a person's health status for the point in time that was not available in the general EHRs.

In this study, our research only focuses on GHE records. However, it can be foreseen that powerful clinical decision support systems will be available in the future when GHE data and EHR patient data can be collected, linked, and considered together.

1.2.3 Formal Definition

After introducing the GHE dataset and discussing its key characteristics, now we can formally define GHE records of a person. A person may have participated in a GHE program for many but possibly non-consecutive years. We call the sequence of a person's GHE records their GHE Record Sequence. Formally, it can be defined as follows.

Definition 1.1 (GHE Record Sequence)**.** *A participant p_i's GHE records are represented as a time-stamped sequence r_i of length n_i, where each element at position j is a tuple (t_{ij}, v_{ij}), where v_{ij} is a d-dimensional mixed variable type vector that stores the GHE results of year t_{ij}. Mathematically, $r_i = ((t_{i1}, v_{i1}), (t_{i2}, v_{i2}), ..., (t_{in_i}, v_{in_i}))$. In addition, we use $r_i(m)$ to denote p_i's records upto the m^{th} record, i.e., $r_i(m) = ((t_{i1}, v_{i1}), (t_{i2}, v_{i2}), ..., (t_{im}, v_{im}))$.*

1.3 Research Challenges

The problem that we address in this thesis is a classification problem on healthcare data. Most work including the one in this thesis focused on health risk prediction based on mining large healthcare data collected from multiple sources, namely the linked GHE and COD datasets. We sometimes refer to the problem as a health score prediction problem in this study, particularly as the Personal Health Index prediction problem, where the score is defined on health risks. The problem is formulated as a classification-based risk prediction problem using the linked main Cause of Death (COD) information as label.

In addition, we also explored another data source, namely a medical literature database. In this classification task, instead of predicting health risks, we rank drugs based on medical literature for drug recommendations.

The key challenges of the research presented in this thesis are closely related to the data characteristics identified in Section 1.2.2, which are listed as follows.

- **Feature Extraction:** GHE record sequences are infrequent, incomplete, sparse, and heterogeneous, as discussed in Section 1.2.2. How to pre-process and extract features from GHE record sequences, in order to capture enough information for effectively representing the health status of a person is a challenge. The considerations include how do we model the time aspect of the GHE record sequences that demonstrates the changes of a person's health over time. This issue is mainly handled in Chapter 3.

- **Heterogeneity:** GHE generally covers a wide range of health assessments, including lab tests, physical examinations, cognitive and mental assessments, as well as personal habits such as exercise and diet habits. These assessments may be recorded in possibly different measurements, scales, and data formats, focusing on different parts of the body. Designing a model that is capable of capturing and harnessing this heterogeneity is challenging. This issue is mainly handled in Chapter 5.

- **Label Uncertainty:** Due to the possible comorbidity and the fact that only one cause would be chosen by an authorized person (e.g., a general practitioner) as the main cause of death in

documenting a death certificate, there is always a certain degree of uncertainty pertaining to these COD labels. Therefore, the challenge becomes how do we build a model that is able to handle label uncertainty. This challenge is handled in Chapter 4.

- **Large Unlabeled Data:** Most GHE participants are without a COD label. In fact, these people account for 92.6% of the GHE participants in our dataset (see Section 1.2.2). However, the health status of these unlabeled GHE participants can vary from healthy to ill. In addition, there is no "ground truth" available for the healthy examples. Therefore, we were left with a machine learning problem with large unlabeled data and with the absence of negative (healthy) cases. Learning a model that is capable of dealing with this issue is very challenging. This problem will be handled in Chapter 5.

1.4 Research Goals

This section discusses the research goals of this thesis. At the highest level, our aim was to build effective classification models for healthcare decision support. An overview of our research goals can be found in Figure 1.2.

The main focus is on Classification-based Health Risk / Score Prediction. This is to enable identification of individuals at risk based on their health examination records. The predicted outputs are expressed as scores, either in terms of risk scores or health scores that we called Personal Health Index (PHI). We first defined as a complement probability of overall health risk and later extended it to a vector of scores representing the health risks in disease categories.

The second focus is on Classification-based Recommendation. This research was to explore additional data sources with the hope to integrate them into our classification-based health risk prediction framework and improve the overall performance. This exploration, particularly in medical literature data, led to a classification-based recommendation problem that is addressed as our fourth goal.

A summary of the goals of our research is listed as follows.

- The primary goal is to design and implement a framework for classification-based prediction as

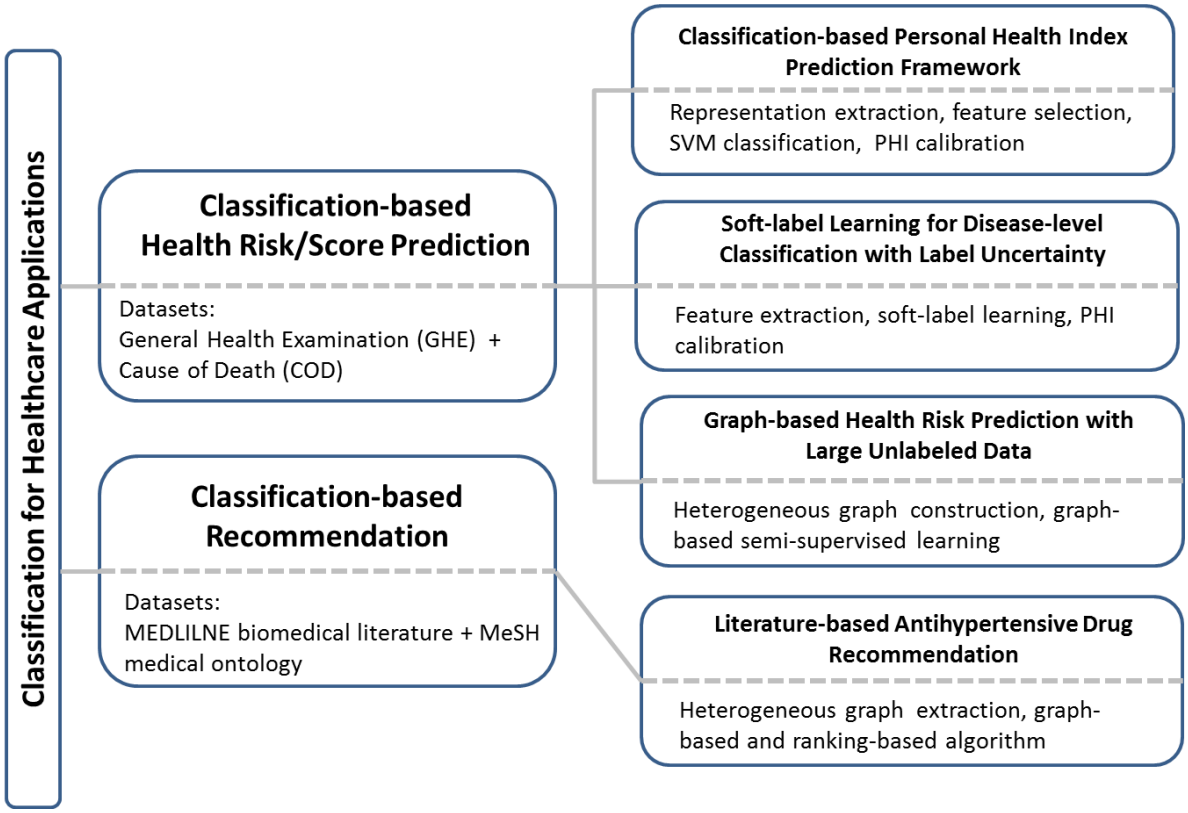

FIGURE 1.2: *Overview of the research goals.*

the first step of the research. Particularly, our approach needed to consider the special characteristics of GHE record sequences, namely irregularity, incompleteness, sparsity, as well as the time aspect or the longitudinality for effective feature extraction and data modeling.

- The second goal was to incorporate the handling of label uncertainty into the classification framework, extending it from an overall prediction to a disease-level prediction. To tackle the label uncertainty issue, we focused on a soft-label learning approach that learns soft labels as health scores for disease categories.

- The third goal was to design a framework that was capable of dealing with large amount of unlabeled data and in the absence of the ground truth examples for "healthy" cases. We proposed a graph-based semi-supervised learning approach to tackle the problem. In order to demonstrate the stability of our proposed algorithm, we evaluate the performance of the proposed algorithm, using both synthetic and real datasets.

- The last goal was to explore medical literature as an additional data source for literature-based

decision support. Particularly, we approached a drug recommendation problem by adapting a graph-based and ranking-based classification algorithm.

1.5 Main Contributions

The main contributions of this thesis are described as follows.

1.5.1 A Classification-based Framework for Personal Health Index Prediction

As the first step of our research, the purpose of this work was to develop and implement a basic framework for health risk prediction. Since the highest health risk is health-related death, we explored the health-related main Cause of Death (COD) information linked to the GHE participants as the labels. Particularly, the COD information was used as binary labels (alive/deceased) and the output predicted score for an individual can be viewed as based on the mortality risk.

In this work, we introduced a health scoring system called Personal Health Index (PHI) to quantify the overall health status of a person. PHI is defined as a complement probability, i.e., $1 -$ mortality risk. The issues we encountered in this study include the following

- **Data pre-processing.** This is a known challenge for medical data modeling. As reported by Esfandiary *et al.* in [37], pre-processing takes about 60% to 90% of the project effort in practice. Key data pre-processing issues encountered in this work included data cleaning, natural language processing on the data collected from free-text entries, sparse data modeling, and missing value handling.

- **Feature extraction.** As described in Section 1.2.1, our GHE dataset had 230 attributes and contained a mixture of numerical, categorical and free-text data types. How we extracted features that preserved the relevant information for effective prediction is critical.

- **Classification and model selection.** The selection of classification models is very often problem-specific and data-driven. The state-of-the-art data mining algorithms were considered and experimentally evaluated for the effectiveness of our framework.

- **Class imbalance problem.** As mentioned earlier in Section 1.2.1, about 7.4% of the participants in GHE dataset were deceased. In other words, the positive class only accounts for 7.4% cases, which makes the class distribution imbalanced.

Our main contributions are listed as follows.

- This work pioneered in applying data mining techniques to the problem of predicting personal health scores based on health examination records of the whole population. Particularly, we defined the Personal Health Index (PHI) to quantify health status and develop a classification-based prediction framework to effectively infer a person's PHI.

- For feature extraction, we proposed representation extraction strategies to handle the longitudinality of the GHE record sequences.

- The effectiveness and stability of the proposed framework have been demonstrated by experiments conducted based on the real health datasets, namely the GHE and COD datasets.

1.5.2 A Soft-label Learning Approach for Disease-level Personal Health Index Prediction with Label Uncertainty

The aim of this work is to extend the basic framework of Personal Health Index (PHI) prediction from predicting an overall health score to a disease-level prediction. In the basic framework, PHI represents a score of overall health, calculated based on the mortality risk. In this work, PHI was redefined as a vector of scores, each reflecting the health risk in a particular disease category. We made use of the disease-level cause of death information from the COD dataset as the classification label.

However, the disease-level COD code suffers from the label uncertainty issue. A person only has one COD code recorded as the main cause of death, but he might suffer from comorbidity. Therefore, his or her COD code only captures part of the real health status. Given such incomplete information, we addressed this as a label uncertainty issue and proposed a soft-label learning solution for the problem. The accurate soft labels are interpreted as health scores.

The main contributions of this work are summarized as follows.

- We extended the concept of PHI from a single overall health score based on all-cause mortality to a vector of scores, each reflecting personal health risk in a disease category.

- Rather than treating labels as 100% certain, as in our previous work, we designed a soft-label learning solution to handle label uncertainty.

- Experimentally we demonstrated the effectiveness of our method based on a large GHE dataset of 262,424 records from 102,258 participants.

1.5.3 A Graph-based Classification Approach for Health Risk Prediction with Large Unlabeled Data

Large unlabeled data is a reality of most healthcare applications. As pointed out in Section 1.2.1, about 92.6% of the participants in our GHE dataset have no COD labels. In addition, their health conditions can vary from generally healthy to very ill. In order to handle this issue of large unlabeled data with no ground truth for negative (healthy) cases, we proposed a graph-based semi-supervised learning approach called SHG-Health that learns soft labels as health scores for a range of disease categories. The proposed algorithm is able to predict an additional "unknown" class to handle the unlabeled cases.

In comparison, our previous work summarised in Section 1.5.2 did not consider the large unlabeled data issue; rather, it randomly selected the unlabeled data the same size as the labeled data and treated these "alive" cases as the "negative" examples. The proposed model in this work takes the reality of large unlabeled data into account. It is able to handle heterogeneity of GHE records by exploring the graph structure. A additional mechanism of learning an extra class is designed for handling the unlabeled data.

The main contributions of this work are listed as follows:

- We presented the SHG-Health algorithm to handle a challenging multi-class classification problem with substantial unlabeled cases which may or may not belong to the known classes. This work pioneered risk prediction based on healthcare data in the presence of large unlabeled data.

- A novel graph extraction mechanism was introduced for handling heterogeneity found in longitudinal health examination records.

- The proposed graph-based semi-supervised learning algorithm SHG-Health combines the advantages from heterogeneous graph learning and class discovery. It demonstrated significant performance gain on a large and comprehensive real health examination dataset as well as synthetic datasets.

1.5.4 A Graph-based Ranking Approach for Literature-based Drug Recommendations

Finally, we diverted ourselves from the GHE and COD datasets to explore another type of data source, namely medical literature, with the hope to eventually integrate the findings into our risk prediction framework. In this study, we approached an anti-hypertensive drug recommendation problem based on the MEDLINE bibliographic database and its linked medical ontology MeSH.

The drug recommendation problem was formulated as a ranking-based classification problem. We adapted our previous work MedRank [20], a graph-based ranking algorithm to solve the problem. To evaluate the system recommended list, we designed a questionnaire to collect drug usage preference from clinical professionals. This work was carried out in my 3-month internship at the Hunan Research Center for Hypertension, Third Xiangya Hospital, Central South University, Changsha, China.

1.6 Thesis Organization

The remainder of the thesis is organized as follows. In Chapter 2, we review the literature related to the research topics in this thesis. Chapter 3 focuses on a classification-based framework for Personal Health Index (PHI) prediction. We introduce the concept of PHI and present a group of representation extraction strategies for effective classification. In Chapter 4, we extend PHI to a vector of disease-level health scores, and propose a soft-label learning approach to handle label uncertainty. Chapter 5 focuses on the issue of large unlabeled data commonly found in real-world health datasets and presents a graph-based semi-supervised learning method as our solution to the problem. In Chapter 6

we adapted a ranking-based classification algorithm to solve a literature-based drug recommendation problem. Finally, the conclusions and future research directions are summarized in Chapter 7.

Chapter 2

Literature Review on Healthcare Data Analysis and Mining

The advances in storage technology, computational power, and the gradual digitalization of healthcare data in the past ten years have led to a new era of healthcare data analysis and modeling [35]. What was once done manually and limited by human capacity can now be assisted by computers on a scale that could hardly have been imagined in the past.

With the overwhelmingly large healthcare data available, the urgent question now is how to design smarter tools to explore, analyze, utilize, and harness the data. In board terms, the field of research is called Health Informatics, which is a discipline that combines information science and computer science within the realm of healthcare with a goal to ultimately improve the quality of care [56]. According to the review of Herland *et al.* [56], Health Informatics can be categorized into five sub-fields, namely Bioinformatics, Neuroinformatics, Clinical Informatics, Public Health Informatics, and Transitional BioInformatics (TBI), based on the level of human existence that they are focusing on (i.e., molecular-level, tissue-level, patient-level, population-level, and combined-level, respectively). Across these sub-fields, data mining is a key tool for big data analytics. In fact, an increasing number of data mining applications has been developed to improve healthcare services over the years [34, 110, 62].

In this thesis, we target classification problems in healthcare applications, particularly classification-based health prediction problems. We will first review the traditional approaches for health scoring

and health data analytics in medical research. Secondly, we will review studies on mining healthcare data especially with a focus on classification. Particularly, we will look at existing work on classification-based risk prediction. In addition, we will review the existing studies related to the challenges identified from our real datasets in Section 1.3, especially, label uncertainty, and large unlabeled data. Finally, we will look at existing work on mining medical literature.

2.1 Traditional Approaches in Medical Research

In medical research, such as biomedicine and public health, it is common to have repeated observations of multiple outcomes [14], resulting in multivariate longitudinal data. Statistical methods are often used for data analysis. The goal for such analysis is to model the multiple responses over time and understand their relationships, especially the correlation. In this section, we will review the existing studies in this domain from two aspects. First, we will summarize existing medical studies on health examination data and their approaches. Secondly, we will discuss studies on health scoring systems developed for clinical decision support.

2.1.1 Data Analytics on Health Examination Data

Medical studies on health examination data analysis are mainly found in the field of epidemiology, particularly in population-based analysis on prevalence and risk factors [108, 41, 126]. A general practice is to formulate a hypothesis on a set of risk factors and validate the hypothesis using a version of regression analysis. The goal of learning model parameters, in this case, is not so much for predicting future cases, as for factor-importance weighting, which is then validated via statistical testing [41, 126].

The datasets in use are mostly based on sampled participants of health examination surveys; therefore, often the longitudinal records of participants are not available. For example, the U.S. National Health and Nutrition Examination Survey (NHANES) datasets [8] are probably the most popular and highly researched health examination datasets. NHANES datasets contain samples about 5,000 participants each year in the U.S. based on certain subgroup proportions predetermined by the authorities. At the time of writing, there were 1,860,000 retrieved publications on NHANES data analysis from

Google Scholar, mostly in the fields of Medicine or Public Health.

The strength of NHANES datasets lies in its data collection method that gives representative samples of the whole population or certain targeted groups. However, except for a small number of follow-ups on certain test items, a person's longitudinal records are generally not available from the NHANES datasets. There are also private datasets with traceable patient history, but they are generally difficult to obtain. For example, a dataset of 462,293 participants was studied for risk factor analysis for all-cause mortality attributable to kidney disease [126].

2.1.2 Health Scoring Systems

In medical practices, it is important to quantify a person's health status into numerical values from observations. Methods that are developed to calculate and attribute a score to an aspect of health to support assessments are generally known as health index or scoring systems. This section will review existing studies in these categories.

Health index

Health indices provide numerical expressions of a person's health status [69]. Such indices may be expressed as a single value aggregated from a list of factors [135, 93], or in terms of a profile of interrelated scores [40]. Generally, "Health Index" refers to a numerical indication of health of a given population, while "Health Status Index" denotes a set of measurements for members of a population [105].

For example, the Charlson's Index [19] was introduced in 1987 for classifying clinical comorbidity in order to assess whether a patient will live long enough to benefit from a specific screening measure or medical intervention. Later, Deyo *et al.* [33] adapted the Charlson's Index for ICD-9-CM administrative databases. Newman *et al.* [93] designed a physiological index for comorbidity based on five noninvasive tests across the full range of values. These type of indices are typically defined based on factors manually selected by experts and validated by epidemiological studies.

For indices designed for personal health status, Yi *et al.* [135] developed a bio-mark based system to grade personal health status. They also call it personal health index. The index is calculated based

on a linear combination of four indices, whose weights are determined by expert knowledge. These four indices are cardiovascular index, stress index, obesity index, and management index. However, their approach has the following three disadvantages: First, they do not consider the multiple years of data available for the same person, which shows a continuous development of a personal health status. Second, they do not take the ultimate health risk (i.e., the death) of a person into account for predicting a person's health status. Third, their computation does not rank an individual's situation against the whole population.

To sum up, the existing approaches of health index computations are purely based on expert knowledge and thus only consider a limited set of factors.

Scoring Systems

Prognostication has long been a common practice for physicians. In order to minimize human subjectivity and errors due to optimism and fatigue, different scoring systems have been introduced. For example, Greenberg *et al.* [50] proposed a prognostic scoring system for myelodysplastic syndromes, known as the International Prognostic Scoring System (IPSS). Many scoring systems that quantify the severity of illness were introduced over the years for places such as Intensive Care Unit (ICU) [43] and palliative care [49]. The most commonly used ones in such intensive care are the Acute Physiology Age and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), and Mortality Prediction Model (MPM). They have been improved over recent years and the latest versions are APACHE IV [145], SAPS 3 [91], and MPM III [58].

Studies have been conducted to compare the discrimination and calibration of these scoring systems in terms of predicting ICU mortality, as well as the effect of the presence and absence of certain variables. For example, Keegan *et al.* [71] showed that APACHE III and APACHE IV had better discrimination than SAPS 3 and MPM III. Moreover, the addition of the resuscitation status variable that is exclusive to MPM III did not improve the prediction of APACHE III, APACHE IV, and SAPS 3. Generally, these methods are defined based on factors selected with expert knowledge and validated via population-based studies [128]. However, as discussed earlier, it becomes problematic when the dimensionality increases and the longitudinal aspect is involved.

Roothman $et\ al.$ filed a patent for a monitoring system that generates patient health score for hospital patient care and management [111]; however, the underlying scoring mechanism is unknown. Online scoring systems for personal health management have also been gaining in popularity in recent years. For example, Nuffield Health Score [5] computes user's health score online based on a set of predefined measurements uploaded from mobile sensor devices. However, neither of them is adaptive to the rapid growth of information, for example collected from health examination records or electronic health records, as the number of factors to be considered in these cases would also grow rapidly.

2.2 Data Mining on Healthcare Data

Data mining is a practice that extracts useful and interesting information from big data [117]. Key Data Mining tasks include classification, regression, clustering, association rule mining and hybrid approach.

With the increasing amount of healthcare data stored digitally over the years, more and more researchers in the Data Mining community have devoted themselves to developing applications to improve healthcare services. According to the classification of Esfandiary *et al.* in their comprehensive review [37], data mining applications in medicine can be viewed in terms of six key activities in medicine, namely, screening, diagnosis, treatment, prognosis, monitoring, and management.

In this section, we will focus our discussions on the classification task. We will start with a general review of existing studies on healthcare data classification, followed by a closer look at the literature on health risk prediction. Then we will survey two special issues in the literature, namely label uncertainty and large unlabeled data.

2.2.1 Classification on Healthcare Data

Classification is a data mining task of assigning objects to one of several predefined categories [117]. In the literature of healthcare data mining, one strain is in morbidity prediction where the emphasis is on predicting specific diseases. Cancer has been the most researched illness amongst all [37, 125, 102, 137]. Studies on other diseases have also been reported, such as but not limited to, diabetes [92, 113],

heart failure [106], mental health problems like depression or suicide [120, 121], or special diseases [96, 90]. Another strain focuses on mortality/survivability prediction, for example, on Intensive Care Unit (ICU) patients [6] or cancer patients [73].

In this section, we will review some representative methods in recent years, with a focus on studies on classification-based risk prediction. We will leave the existing works with a special focus on label uncertainty and unlabeled data issues to Section 2.2.2 and Section 2.2.3

Wei *et al.* [125] proposed an ensemble method for cancer prediction on gene expression data. They adapted a hypercuboid learning approach based on rough set theory to provide more interpretable prediction. Nguyen *et al.* [96] introduced a classification method for *heparin induced thrombocytopenia*. Their model took into account the soft-label information that reflected how strongly the human expert felt about the original class labels. Specifically, they added an auxiliary term for soft label in the standard SVM formulation and showed improved performance on their dataset.

Pan *et al.* [102] proposed a boosting approach for graph-based cost-sensitive classification called CogBoost to handle network data such as chemical compounds for anti-cancer activity prediction. Their iterative algorithm involves a depth-first search optimal subgraph exploration step, a risk minimization step with linear programming, and a graph weight update step. CogBoost outperforms the other boosting-based and SVM-based methods empirically.

To deal with multivariate time series that are often found in ECG and ICU data, Ghalwash *et al.* [47] formulated a logistic-loss constraint optimization problem for interpretable multivariate pattern extraction. The effectiveness of the proposed method has been demonstrated in terms of classification performance based on an ECG dataset and two blood gene expression datasets, although the datasets they used were rather small (i.e., 17, 20, 448 patients respectively).

In the following subsection, we will review classification-based studies that have a focus on predicting health risks.

Classification for Health Risk Prediction

Most existing studies on medical risk classification are based on patient EHRs. Commonly studied diseases for screening data classification are cancer, Alzheimer's disease, and diabetes [37]. Mohktar *et al.* introduced an exacerbation risk prediction system for chronic obstructive pulmonary disease

using home tele-health measurement data [90]. They used a classic classifier, namely the CART algorithm.

Liu *et al.* [85] introduced a scoring system for cardiac arrest prediction based on the ECG signals, vital signs, and demographic data of 1,386 critically ill patients. The initial score was calculated based on geometric distances between the query point and the cluster centres of the positive class and the negative class respectively. Then the score was updated using SVM-based classification and neighborhood checking. They also adopted feature extraction methods, such as Principal Component Analysis (PCA), Kernel PCA, and Linear Discriminant Analysis (LDA). Simon *et al.* [113] developed a method for early diabetes mellitus risk assessment based on association rule summarisation techniques. The area they targeted was the large amount of rules returned by traditional association rule mining algorithms.

To handle data sparsity, Wang *et al.* [122] proposed a multilinear sparse logistic regression model for clinical risk prediction to handle especially EHRs that come in a tensor form, for example, the functional MRI (fMRI) data. The method has been verified using a synthetic dataset and a fMRI dataset of 1,005 Alzheimer's Disease patients. Their proposed algorithm was shown to outperform classical classifiers such as Nearest Neighbor and Support Vector Machines, as well as other forms of Logistic Regression models such as sparse logistic regression. However, their method assumes that the input tensors have a fixed dimensionality, so it is not directly applicable to longitudinal EHRs with various lengths.

Qian *et al.* [106] designed a relative-similarity-based method for active patient risk prediction. They proposed a linear neighborhood propagation framework to learn a patient similarity matrix, based on which patient risk can be predicted. Their method was evaluated using UCI breast cancer and diabetes data, MRI data of patients with Alzheimers disease, as well as real EHR data of congestive heart failure patients.

In terms of how the longitudinal aspect of EHR is handled, Neuvirth $et\ al.$ [92] proposed a framework for personalized care management based on time-series data from diabetes patients. They compared classical models for binary classification, using different feature selection and extraction strategies. Time-series is merged along time axis by computing the mean on each feature.

Wiens $et\ al.$ [127] introduced a novel concept of risk time-series. In their two-stage SVM-based

framework for hospital acquired *Clostridium difficile* risk classification, multi-dimensional time-series patient records are reduced to one-dimensional risk time-series, from which summary statistics were extracted for classification. Tran $et\ al.$ [120] introduced a suicide risk classification framework based on a dataset of 10,000 patients dataset. They modelled longitudinal patient records as an image, and applied a bank of one-side filters for temporal feature extraction.

Summary

As a summary, most existing works use patient data, especially patient EHRs, collected from people who were already suspected or diagnosed to be sick [121, 90, 75, 96, 102, 141]. However, mining general health examination data is an area that has not yet been well-explored, except a few studies on risk prediction such as the chronic disease early warning system proposed in [64].

In [64], Jen *et al.* extracted 53 features from medical examination data and case histories of 2,724 cases of six chronic diseases. They proposed a classification framework based on Linear Discriminant Analysis (LDA) for feature extraction and the K Nearest Neighbor (KNN) classifier.

However, their methodology was not comparable, because they only used positive cases (patients of pre-selected 6 diseases) and did not concern the potentially harder-to-handle negative cases (i.e., those without disease labels). Their goal was a multi-class classification problem, namely to correctly classify the disease that a given patient would have amongst six possible diseases.

2.2.2 Classification with Label Uncertainty

Label uncertainty is often treated as label noise in the literature. Label noise may come from four different sources according to the survey of Frenay *et al.* [44], namely human expert inadequacy (e.g., unknown testing result), non-expert unreliability (e.g., crowd sourcing), human subjectivity (e.g., medical judgement), and data encoding or communication problems.

In the medical domain, key sources of label noise are expert subjectivity and inadequate information in clinical judgments such as diagnosis [104]. Such noise is usually regarded as mislabeling to be detected and corrected [87, 107, 45]. For example, Garca-Zattera *et al.* employed binary Markov

models to estimate misclassification parameters for dental research [45]. Rantalainen *et al.* introduced a Bayesian approach to detect control subjects who might be actually the undiagnosed cases in a case-control study [107]. The underlying assumption of these approaches was that ground-truth labels exist and those labels are certain, so the task was to detect and correct mislabeling caused by human inadequacy.

However, uncertainty may come from multiple, non-exclusive medical conditions [82]. For example, health-related death can be caused by the complications of several co-existing diseases, and the identified main cause of death (COD) may only explain the death to a certain degree. In such cases, rather than identifying "mislabeled" cases, it does make sense to allow labels to have a certain degree of certainties.

Lin *et al.* introduced a decision support system for lower back pain diagnosis that expresses diagnostic uncertainty using verbal probability estimation; however, their system was highly expert-dependent as the knowledge base relied on experts to maintain the updates. Recently Nguyen *et al.* [96] introduced a learning approach that considered the auxiliary soft-label information collected from human experts to quantify label uncertainty. They included an additional term in the original SVM formulation to define the loss of not respecting the orders induced by the auxiliary subjective probabilities. Their proposed methods showed better performance than methods that did not consider the auxiliary soft-label information.

In the field of Computer Vision, Yi *et al.* [133] introduced a soft-label learning approach for complex event detection on Web videos. In the context of event detection on videos, there were often only a few positive instances available for training. On the other hand, there were instances where the relationship to the target event was uncertain. They formulated the learning problem as a soft-label optimization problem and demonstrated that exploiting related examples improved performance. Their work formed a base for our work on handling the label uncertainty issue in health examination data presented in Chapter 4.

2.2.3 Classification with Unlabeled Healthcare Data

In this section, we review key approaches that consider unlabeled healthcare data, namely the Positive and Unlabeled Learning approach, the Semi-Supervised Learning approach, and the Active Learning approach.

Positive and Unlabeled Learning Approaches

Methods proposed for handling and utilizing large unlabeled healthcare data are mostly found in a closely related field of disease gene classification. One common approach is *Positive and Unlabeled (PU) learning* [83] that learns from positive and unlabeled data alone. It is used when only the labels for disease genes are available [130, 131]. For example, Yang *et al.* introduced an ensemble-based PU learning approach for robust disease gene prediction.

Recently, Nguyen *et al.* proposed *mPUL* [95], a multi-class PU learning model for activity recognition. The method trains m 1-vs-others binary probabilistic base classifiers, each trained with a positive set and a merged set of negative and unlabeled instances. The class decision was based on the maximum class probability greater than 0.5; otherwise the unknown class would be predicted. However, the method did not consider the heterogeneous nature of the examination records. In addition, their method used a combined set of negative and unlabeled example, while in our case negative examples are not available, because there is no ground truth for "healthy" participants.

Semi-Supervised Learning Approaches

Semi-Supervised Learning [144] (SSL) that learns from both labeled and unlabeled data has attracted increasing attention in healthcare applications based on EHRs [73, 60, 97, 46, 124, 61, 84, 138]. PU learning can be seen as a special case of SSL.

At the molecular level, Huang *et al.* proposed iSELF [60], a SSL method based on improved local Fisher discrimination analysis for disease gene classification. Their method was parameter free and was claimed to be statistically uncorrelated by imposing an uncorrelated constraint on the objective function. Nguyen *et al.* [97] constructed a protein-protein interaction network, which defines interacted genes as candidate genes and the rest as negative genes for SSL based on Gaussian fields and

harmonic functions [144]. In this work, they defined eight scores functions to extract topological and genomic/proteomic features from six biological databases. The experimental results showed that the proposed SSL-based method outperforms SVM and KNN classifiers.

At the disease level, many graph-based SSL methods were proposed. Garla *et al.* [46] applied Laplacian SVM as a SSL approach for cancer case management based on clinical text data. They extracted named entities from clinical text and represented each document using a bag-of-word representation, which resulted in feature vectors with more than 15,000 dimensions. Their experiments showed that the Laplacian SVM classifier outperformed the traditional SVM classifier.

Wang *et al.* [124] proposed a graph-based SSL method that was able to learn patient risk groups for patient risk stratification. Based on the observation that risk factors and their importance may vary across different patient groups, they added in their objective function a grouping term that assigned patients into risk groups and a matching term that indicated the group assignment and label prediction. Their proposed method outperformed classical SSL method such as GGSSL [99] on UCI datasets. They also demonstrated six risk groups identified by their method from a dataset with 1,296 Congestive Heart Failure (CHF) patients.

Kim *et al.* [73] proposed a co-training graph-based SSL method for breast cancer survivability prediction that utilized the so-called pseudo-labels. The algorithm iteratively assigns pseudo-labels to unlabeled data when there was a consensus amongst the learners. In the next iteration, it includes the pseudo-labeled instances in the labeled set for training. The iteration terminates when the unlabeled set stops decreasing. Their method has better classification accuracy than ANN, SVM, and standard SSL classifiers on most of the datasets. Zhang *et al.* [138] proposed a ranking-based approach for lung nodule image classification. Their algorithm first constructed a bipartite graph that captures the relationship between the labeled and unlabeled instances. Based on the constructed bipartite graph, the similarity scores were computed for the purpose of classification.

To capture the time aspect of the data, Liu *et al.* [84] constructed a temporal graph based on medical event sequences of potential patients for temporal phenotyping. They formulated the learning problem as a matrix reconstruction problem with a goal to minimize the reconstruction error. The algorithm learned graph bases as phenotypes and their weights. Their proposed temporal phenotyping approach has been shown to outperform aggregated vector representation and other bag-of-pattern

representation for classification.

However, different from our case, none of these methods considered an "unknown" class and they all have predefined instances for all classes, either by experts [60, 46, 124, 73, 138, 84] or via other mechanisms [97]. In addition, unlike our approach, all the mentioned graph-based SSL methods used homogeneous graphs. Moreover, each patient was represented by a set of International Classification of Diseases (ICD) codes in [124], which is very different from a sequence of health examination records in our problem setting.

Active Learning

Another field of study that addresses the unlabeled data issue is Active Learning [29], where the concern is how to design a learner that is able to actively acquire data instances from the training set. Generally speaking, a learning algorithm iteratively selects currently unlabeled instances and queries an oracle to obtain their labels [16]. It is especially useful in application domains where the amount of unlabeled data is huge and the availability of oracle, particularly a human expert, is limited.

Active learning has been applied to EHR data predictions [106, 24]. Chen *et al.* [24] integrated an uncertainty sampling active learning approach with SVM-based phenotyping algorithms for three types of diseases, namely rheumatoid arthritis (RA), colorectal cancer (CRC), and venous thromboembolism (VTE). The algorithm starts with training an initial SVM classifier with two samples and the trained model predicts the labels for the rest of the training samples. Then an uncertain sampling method is used to query a sample closest to the decision hyperplane. The sample is annotated and included to retrain the SVM learner. This process is repeated until the unlabeled pool is empty.

Qian *et al.* [106] proposed a query selection strategy that, instead of asking absolute questions such as labels, the algorithm asked the human experts to answer the relative similarity of the neighboring points to the queried point by placing ordering on them. They formulated the problem as a quadratic optimization problem that minimizes the reconstruction error of a patient similarity matrix. The ordering provided by the expert was encoded as the constraints in the optimization problem.

Although active learning is effective with unlabeled data, it is not applicable to our current problem for the following reasons: First, if a human expert was used as the oracle, it was not available in our current setting. The label used in this study is Cause of Death (COD) of the health examination

participants, which would not be obtainable by querying an oracle, but only available when a person has passed away. Second, if the ground truth labels were to be used as an oracle, we would have needed a large amount of labels so that some could be hidden in the training phase for the purpose of evaluation in the testing phase, as most studies on active learning did [16]. But this would require larger amounts of labeled data. However, this is an unpractical requirement for healthcare data because our dataset only had 7.4% cases labeled. In addition, using only the available labeled cases (with some labels hidden) in the experiments would mean limiting the search space of a learner and the real unlabeled data would be left unexplored.

2.3 Text Mining on Medical Literature

The past decade has witnessed dramatic growth in electronic biomedical literature, which is generated and stored digitally and is often regarded as a reliable information source [11].

On the one hand, researches have been developing methods to improve document retrieval from biomedical literature databases [78, 136, 54, 13], such as MEDLINE [3], a bibliographic database of life science literature, maintained by the U.S. National Library of Medicine. Leonard *et al.* [78] proposed a Bayesian approach to mine gene and protein references from MEDLINE abstract and title. Yoo et al. [136] conducted a comprehensive comparison study to verify if MeSH ontology would help document clustering for the purpose of improving document retrieval. Haynes *et al.* [54] compared search strategies for retrieving scientifically strong studies of treatment from MEDLINE. Recently, Avillach *et al.* [13] designed a method to automatically detect known adverse drug reactions in MEDLINE, based on Medical Subject Headings (MeSH) including "descriptors", "supplementary concepts" with subheadings such as "chemically induced", "adverse effects" and "pharmacological action".

On the other hand, more and more studies based on text mining techniques have been developed for extracting useful knowledge from MEDLINE. De Bruijn *et al.* [32] proposed a semi-Markov modeling approach to extract medical concepts such as problem, test, treatment from medical texts. Padmini *et al.* [114] used text mining to discover novel uses for Curcuma longa in 2004. They found that curcumin played a beneficial role in retinal disease, Crohn's disease, and spinal cord disorders by

analyzing the influence of curcumin on related genes.

For a network-based approach, Kilicoglu *et al.* developed SemMedDB [72], a semantic network for literature-based discovery. The network was composed of entities extracted from MEDLINE, each represented as a (subject,predicate,object) triple. Kastrin *et al.* [70] proposed a link prediction method based on SemMedDB for literature-based discovery. A comprehensive review on literature-based discovery for novel therapeutic approaches can be found in [59].

In our previous work [20], we proposed MedRank, a ranking-based algorithm for treatment recommendation on a heterogeneous graph extracted from MEDLINE using MeSH ontology. Ji *et al.* [67] proposed a graph-based framework to infer strong relevance heterogeneous entities from unstructured text. They constructed a heterogeneous graph of five types of entities, namely *Drug*, *Disease*, *Compounds*, *Target* and *MeSH* terms, extracted from MEDLINE literature, with additional domain specific knowledge from external data sources, such as U.S. Food and Drug Administration (FDA) approved drugs, human disease ontology, chemical compounds with drug indicator, a dictionary of targets, as well as the MeSH tree.

Chapter 3

Classification-based Personal Health Index Prediction Framework

3.1 Problems and Challenges

Health status, being an abstract concept, is often understood through indices or scores that quantize aspects of well-being into numerical values. Although proven to be very useful, these indices rely on manual selection of factors based on expert knowledge and validated by epidemiological studies [69]. Less rigorous indices are also available for personal health management. For example, Nuffield Health Score [5] computes a user's health score online based on a set of predefined measurements uploaded from mobile sensor devices. However, neither of them are adaptive to rapid growth of information such as the health examination records, as the number of factors to be considered also growing rapidly.

In this study, we define a new health index called *Personal Health Index* (PHI) focusing on providing feedback to individuals based on the evaluation of risks revealed in their health examination records. PHI provides an overall health score that quantifies a person's health status against the whole population, based on data mining technologies.

Perceiving that health is a concept relative to illness, we adopt an approach of "definition by negation" and assume that a person is healthy if he or she is unlikely to have high health risks. Further, we view the PHI prediction as a classification problem and define the score based on a class

membership probability.

Definition 3.1 (PHI)**.** *The PHI of a person p_i is the complement of the probability that p_i is at high health risk.*

The highest health risk is death. We use the main cause of death information as label. The Cause of Death (COD) dataset stores the cases of people who have been taking General Health Examinations (GHE) for some years and then have passed away.

For the detailed description of our GHE and COD datasets, readers are directed to Chapter 1.2.1, where Section 1.1 gives an example of the record sequence of a participant. The key characteristics of both the GHE and COD datasets are also given in that section. The common attribute in both GHE and COD datasets is the de-identified Person ID p_i. Thus the records in COD dataset can be linked with that of the GHE dataset. Excluding the deaths that were caused by unexpected situations such as traffic accidents or crime-related deaths, the health-related death of a person is the highest risk associated with the GHE records.

In this classification task, we only use the COD information as binary labels, where a positive label value (+1) indicates that a person has passed away, while a negative label value (-1) is assigned to a person who is alive, that is, the corresponding record does not have a linked COD label. Note that in this work we regard the alive cases as the negative cases in our binary classification problem. However, in our later work on handling unlabeled data in Chapter 5, those alive cases are treated as unlabeled cases.

Our research problem for mining personal health index is formulated as follows.

Problem 3.1. *Given a set of GHE records of M participants, which are linked to COD dataset with mortality labels $y_i \in \{-1, +1\}$, $i \in \{1, ..., M\}$, predict PHI for each participant.*

3.2 Personal Health Index Prediction Framework

This section describes the proposed PHI prediction framework. As shown in Figure 3.1, the framework contains five main components, namely, *data pre-processing, representation extraction, feature selection, classification, and PHI calibration*. The inputs are a population's archive GHE records

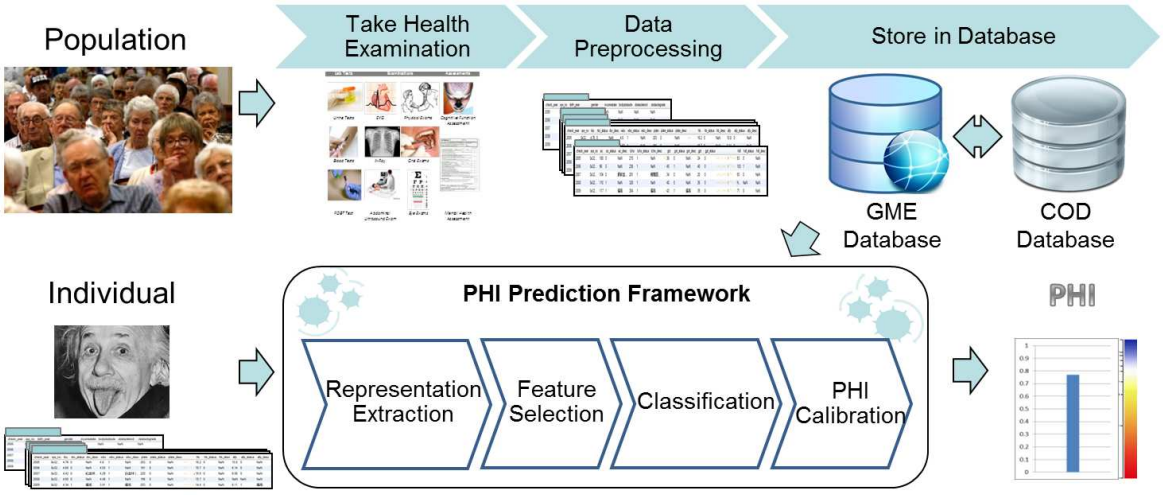

FIGURE 3.1: *An overview of the proposed PHI prediction framework.*

linked to a COD database, and the records of a given participant p_i. The output is the predicted PHI score for the person.

3.2.1 Data Pre-processing

As a real-world dataset, GHE data (Section 1.2.1) contains observations collected from a range of different measurements for the assessments of different parts of the body. This kind of input variety renders GHE data a mixture of value formats, including numerical, ordinal and categorical values, as well as free text entries in Mandarin Chinese. Here we describe how issues were handled in the data-preprocessing stage.

Data Cleaning

As described in Section 1.2.1, there were three fields for recording lab test observations, namely numerical value, status, and descriptions. Although there were fixed ranges of values for *numerical results* and *status* fields, GHE management system does not have strict constraints on data entry at the time of data collection. Because of this, we conducted extensive data cleaning to remove extra or unrecognisable symbols, converted wide characteristics from Asian-based key-in systems into narrow characteristics, and corrected obvious misspellings.

Free Text Entries

Most free-text entry fields, such as *description* for lab test results, doctor suggestions, and patient-reported disease history, were recorded in plain Chinese text. These fields contained rich information yet numerous misspellings and synonyms.

Processing this text would require separate research for natural language processing in Chinese and a comprehensive Chinese medical ontology. We selected fields with less noise, namely, the "reasons for taking medicine" field. The top seven most frequently reported reasons were extracted. They were recorded as seven binary attributes, where a value "1" indicates the occurrence of a reason while a value "0" indicates the absence of that reason.

Missing Values

Missing value is a common issue in medical data. "Missing-ness" observed in our dataset presents mainly in the following three ways:

- Firstly, a small portion of individual nulls is randomly scattered in the records.
- Secondly, certain test items only appear in the GHE program for some years. So their values are missing in the records of other years.
- Thirdly, a person may not participate every year, due to the voluntary nature of the program, so a whole year's records may be missing.

The average number of check-ups per person in our data was 2.56, which indicates that this type of "missing-ness" has an impact on the shape of the data. More precisely, if we view a participant's records as a high-dimensional longitudinal series, it is likely to be rather short and incomplete. For this reason, missing value handling is considered together with representation extraction described in Section 3.2.2.

Binarization

The details of GHE attributes and value handling are discussed in this subsection. Any discrete attribute obtained according to the process discussed here is then further binarized if they were not

already in the binary form.

For patient demographics, age was firstly discretized into five-year bins, i.e., [65, 70), [70, 75), [75, 80), [80, 85), and [85+, where the rounded brackets indicate exclusive interval and the square brackets indicate inclusive interval.

For patient habits, most attributes are binary except reason-for-taking-medicine which is a patient reported field. The reasons with top frequency count were extracted and represented as seven binary attributes as described previously in this section.

For lab tests and physical examinations, there are three fields to record the results, namely observed value, status, and description. The observed values are generally numeric. The status fields indicate whether or not the result of a test is normal. Their values can be either binary or ordinal, depending on the type of tests. The descriptions are in free text format. We only used the information from the status fields for the following reasons.

- Firstly, the reference ranges of these items may differ among hospitals and the information regarding where an examination was taken is not available in the dataset for privacy reasons.
- Secondly, the values for the description fields are mostly missing.

The results of five Brief Symptom Rating Scale (BSRS) questions for mental health assessments were encoded as ordinal values, in terms of degree of severity, i.e., normal, mild, medium, and severe.

For the Short Portable Mental Status Questionnaire (SPMSQ) cognitive function assessment, the overall result was scored in terms of the number of questions that were incorrectly answered. It is an ordinal attribute with values “sound” (0-5 scores), “mild” (6-9 scores), “medium” (10-14 scores), and “severe” (above 15 scores) according to the standard categorization.

3.2.2 Representation Extraction

Feature extraction is a critical step in our framework, because the performance of a classifier is ultimately determined by the amount of information conveyed in input data. In the literature many transformation algorithms for time-series data were proposed, such as wavelet-based transformation and the Symbolic Aggregate approXimation (SAX) transformation [123]. These methods were usually applied to high-frequency time-series, such as ECG signals recorded in the scale of Hz [109].

Instead of classifying directly on the longitudinal data, we took a feature-based approach [129] that reduces sequence classification into a point-based classification due to the facts.

- Firstly, classical point-based classifiers can be applied.
- Secondly, the missing value issue can be implicitly handled in such reduction.

Four representation strategies are defined to extract patient feature vectors from the GHE record sequences. The first two are baseline methods and last two are our proposed methods, as discussed below in order:

- **Avg** is our first baseline method that averages over a participant p_i's n_i records r_i across time on every dimension separately:

$$Avg(r_i) = 1/n_i \cdot \sum_{j=1}^{n_i} v_{ij}. \tag{3.1}$$

  where $r_i = ((t_{ij}, v_{ij}))_{j \in \{1,\ldots,n_i\}}$.

  Although this approach considers all the records of p_i, it does not capture the longitudinal aspect of a participant's records, because it treats every year equally.

- **Last** is our second baseline method. It uses the last GHE record of the record sequence, i.e., v_{in_i}, to represent a participant p_i's records throughout years:

$$Last(r_i) = v_{in_i} \tag{3.2}$$

  The intuition behind this method is that the latest record best reflects the current status of a person's health.

- **Stacked** is our first proposed method. It was designed as a combination of the baselines, in order to capture their advantages. This approach stacks the *Avg* representation of p_i's all previous records except the last one and the *Last* representation to form a new feature vector:

$$Stacked(r_i) = [Avg(r_i(m)), Last(r_i)] \tag{3.3}$$

This design emphasizes the importance of the latest record, while it still takes the previous records into account. Note that previous records are weighted equally in this approach.

- **Decay** is our second proposed method. It assigns weights to the records in a sequence according to their recency.

  A normalized exponential decay function $K_\alpha(\cdot)$ with a decay scale α is defined as follows:

$$K_\alpha(x) = \begin{cases} \frac{e^{-\alpha x}}{\sum_{t=0}^{H-1} e^{-\alpha x}} & t \in [0, H-1] \\ 0 & \textit{otherwise} \end{cases} \tag{3.4}$$

  where H is the window of the years under consideration.

  Given a participant p_i's j^{th} record (t_{ij}, v_{ij}) of his record sequence $((t_{i1}, v_{i1}), ..., (t_{in_i}, v_{in_i}))$, the weight of (t_{ij}, v_{ij}) is computed by $K_\alpha(t_{in_i} - t_{ij})$, because the direction of decay is from the latest time backwards to capture the recency.

  So, the *Decay* representation is computed by the following function:

$$Decay(r_i) = \sum_{j=1}^{n_i} K_\alpha(t_{in_i} - tij) \cdot v_{ij} \tag{3.5}$$

Should any null values exist after the above extraction process, they are imputed with zero.

3.2.3 Feature Selection

In order to avoid noise from irrelevant features and discover key factors, we further included a feature selection step in the framework. This step was designed especially for non-linear kernels to be discussed in Section 3.2.4, because they do not compute feature weights in the original feature space. Two selection approaches are considered, namely univariate feature selection and L1 regularization.

Univariate F-score Feature Selection

This approach computes feature score independently for each feature. The F score of feature j is computed by Eq. (3.6) as in [25]:

$$F(j) = \frac{(\overline{x_j}^{(+)} - \overline{x_j})^2 + (\overline{x_j}^{(-)} - \overline{x_j})^2}{1/(n_+ - 1)\sum_{i=1}^{n+}(x_{i,j}^{(+)} - \overline{x_j}^{(+)})^2 + 1/(n_- - 1)\sum_{i=1}^{n-}(x_{i,j}^{(-)} - \overline{x_j}^{(-)})^2} \quad (3.6)$$

where $\overline{x_j}$, $\overline{x_j}^{(+)}$ and $\overline{x_j}^{(-)}$ are the average of the jth feature of the whole, positive, and negative datasets, respectively; $x_{i,j}^{(+)}$ is the jth feature of the ith positive instance and $x_{i,j}^{(-)}$ is the jth feature of the ith negative instance.

All features were ordered according to their F scores computed according to Eq. (3.6). Then a subset of features was selected based on a user-defined percentile of top-scored features.

L1 Regularization

L1 regularization uses a penalty term that encourages the sum of the absolute values of parameters to be small, which frequently leads to a sparse parameter vector. This mechanism has been commonly used for feature selection.

For linear classifiers, L1 regularization can be embedded as an extra term in an objective function; however for non-linear models an extra step is needed. Here we use Logistic Regression with L1 regularization [94] as a method for feature selection. Logistic Regression models the conditional probability $P_{\mathbf{w}}(y = \pm 1|\mathbf{x})$. The *prime* problem with the L1-norm regularization is defined:

$$\min_{\mathbf{w}} ||\mathbf{w}||_1 + C\sum_{i=1}^{l} log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}) \quad (3.7)$$

where $\mathbf{w}$ is a weight vector and the meta-parameter C controls the sparsity of $\mathbf{w}$. Features are selected by $\mathbf{w}$.

3.2.4 Classification Models

Recall that the proposed PHI defined earlier in Definition 3.1 was computed based on the class membership of the high risk class. For this classification task, we applied the popular Support Vector Machines (SVM) algorithm [112] with various kernels. SVM finds the decision boundary that maximizes the margin, i.e., the distance between the boundary hyperplane. The nearest class instances that

define the margin are called the support vectors.

To allow possible overlapping class distributions, we adopted a soft-margin SVM with slack variables $\xi_i \geq 0$. Given a training set $(\mathbf{x_i}, y_i), i = 1, ..., l$ where $\mathbf{x}_i \in \Re^D$ and $\mathbf{y} \in \{1, -1\}$, the lost function is defined as:

$$\min_{\mathbf{w},b,\xi} \mathbf{w^T w} + C \sum_{i=1}^{m} \xi_i$$
$$\text{subject to } y_i(\mathbf{w^T}\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$$

where $C > 0$ is the penalty of the error term, and $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\mathbf{T}\phi(\mathbf{x}_j)$ is the kernel function.

Three commonly used kernels, namely the linear (LinSVM), radial basis function (SVM-RBF), and sigmoid kernels (SVM-Sigmoid), are considered in our framework. They are listed below, with kernel parameters γ and r.

- Linear kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = {\mathbf{x}_i}^T \mathbf{x}_j$$

- RBF kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2), \gamma > 0$$

- Sigmoid kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma {\mathbf{x}_i}^T \mathbf{x}_j + r)$$

3.2.5 Personal Health Index Calibration

To further compute PHI based on classification results, we use Platt's scaling [81] to transform the outputs of a classification model into a probability distribution over classes. The Platt's scaling method is defined as follows:

$$Pr(y = 1|\mathbf{x}) \approx P_{A,B}(f) = \frac{1}{1 + e^{Af+B}} \tag{3.8}$$

where f is the decision function of SVM classifier, and A, B are the parameters to be learned.

With this probability of the high mortality risk class ($y = 1$), PHI is computed as:

$$PHI(\mathbf{x}) = 1 - Pr(y = 1|\mathbf{x}) \tag{3.9}$$

By definition, PHI is a value within the range of [0,1].

3.3 Experiments and Results

In this section, we report and analyze our experiment results. Since we are handling record sequences, how these sequences are aligned will be explained first. Then the experiment setup and evaluation measures will be defined. Finally, the analysis of the experimental results and discussion will be given.

3.3.1 Data Alignment

It has been observed in the deceased cases that time gap between a participant's last record and death varies, which implies that the predictability of participants' record sequences also varies. In order to align participant sequences, we selected data subsets by defining the upper bound on the time gap k.

Three data subsets, S_1, S_2, and S_3 were generated by selecting from the original dataset which we refer to as S_0. S_1, S_2, and S_3 contain participants who are deceased/alive within k years ($k = 1, 2, 3$) of their last check-up.

For example, S_1 data subset contains all records of participants who passed away within the first year of their last check-up (positive cases) or those who were alive within the time frame (negative cases). The sizes of the subsets are listed in Table 3.1. If we denote $S_k(d)$ and $S_k(a)$ as the deceased and alive cases of the within-k-year subset, the relationships between these subsets are as follows:

$$S_1(d) \subset S_2(d) \subset S_3(d)$$

and

$$S_3(a) \subset S_2(a) \subset S_1(a)$$

TABLE 3.1: Within k year subset statistics (before resampling)

Within k year subset	Number of participants deceased	%	alive	Number of records deceased	%	alive
$S_1(k=1)$	1,469	2.4	59,021	4,415	2.0	215,462
$S_2(k=2)$	2,433	4.0	58,057	7,100	3.2	212,777
$S_3(k=3)$	2,920	4.8	57,570	8,309	3.7	211,568

where $A \subset B$ means set A is a strict subset of set B.

Note that the percentages of positive cases (deceased) for S_1, S_2, S_3 are 2.4%, 4.0% and 4.8% respectively and 7.4% for the whole dataset S_0. These figures indicate that class distributions of the subsets and the whole dataset are heavily imbalanced. To handle this class imbalance issue, we performed an adaptive re-sampling method called one side selection [12] on each subset. In a binary classification scenario, the method randomly sampled from the majority class the same number of participants as the minority class. For the rest of the chapter, we use S_1, S_2, S_3 to denote the re-sampled within 1, 2, 3 year subsets respectively. Their sizes have $|S1| < |S_2| < |S_3|$ relationships, where $|\cdot|$ denotes the size.

3.3.2 Experiment Setup

We used a Dell workstation with common specification (CPU i7 and 16GB RAM) for the experiments. The experiments were conducted on subsets S_1, S_2, S_3 separately on four representations defined in Section 3.2.2. In addition to SVM classifiers, we included the other two popular classifiers in medical applications, namely K-Nearest Neighbour (KNN) and the CART decision tree model as baselines [7].

In order to fully utilize the data, we conducted 10-fold cross-validation in all experiments. In each fold, three steps are performed as follows.

- **Step 1**: feature selection models were trained and applied to transform training and testing sets.
- **Step 2**: classifiers were trained with the transformed training set using 5-fold cross-validation grid search for the optimal meta-parameter C for all SVM classifiers and γ for SVMs with RBF and Sigmoid kernel. All parameters were tuned in the range of $\{10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}\}$.
- **Step 3**: the learned classifiers were tested with the transformed testing set.

The algorithms were implemented in Python using LibSVM [18] and scikit-learn [7] packages.

3.3.3 Evaluation Measures

For the purpose of evaluating the binary classification results, we adopted five commonly used measures. In order to define these measures, we first define the following notations known as the confusion matrix.

- True Positive (TP): the number of correctly predicted positive cases.
- True Negative (TN): the number of correctly predicted negative cases.
- False Positive (FP): the number of negative cases that was incorrectly predicted as positive cases.
- False Negative (FN): the number of positive cases that was incorrectly predicted as negative cases.

The five measures used for our evaluation are Accuracy, Precision, Recall, F1-score, and Area Under the receiver operator characteristic Curve (AUC), which are defined as follows.

- **Accuracy** computes the portion of correct predicted cases amongst all cases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** calculates the portion of predicated positive class that is actually correct.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall** computes the portion of ground truth class that is correctly predicted.

$$Recall = \frac{TP}{TP + FN}$$

Recall is also known in the medical domain as the sensitivity.

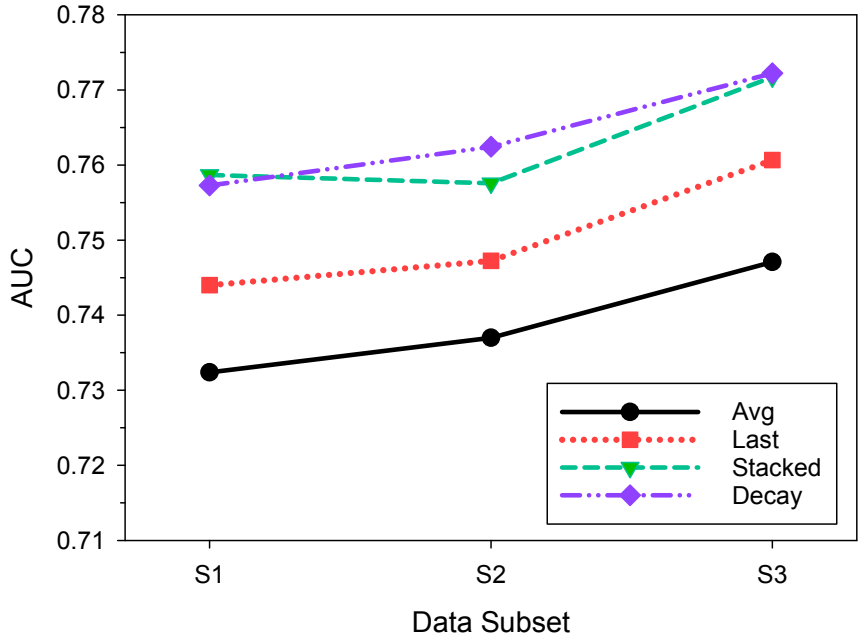

FIGURE 3.2: *Comparing representation extraction methods by classifiers in AUC (LinSVM).*

- **F1-score** is a harmonic mean of precision and recall, defined as:

$$F1Score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

- **AUC** measures the area under the Receiver Operator Characteristic (ROC) curve. Each point on the ROC curve is plotted with the true positive rates or sensitivity and false negative rates or 1-specificity of a given prediction threshold.

3.3.4 Results

In this subsection we analyse the experiment results. The averaged training and testing time with non-linear kernel SVM plus feature selection were 35.21 seconds and 0.07 seconds respectively. The effectiveness of the proposed framework was demonstrated via experimental studies discussed below. As will be shown, the best performance was seen at 77.98% AUC with F-score feature selection.

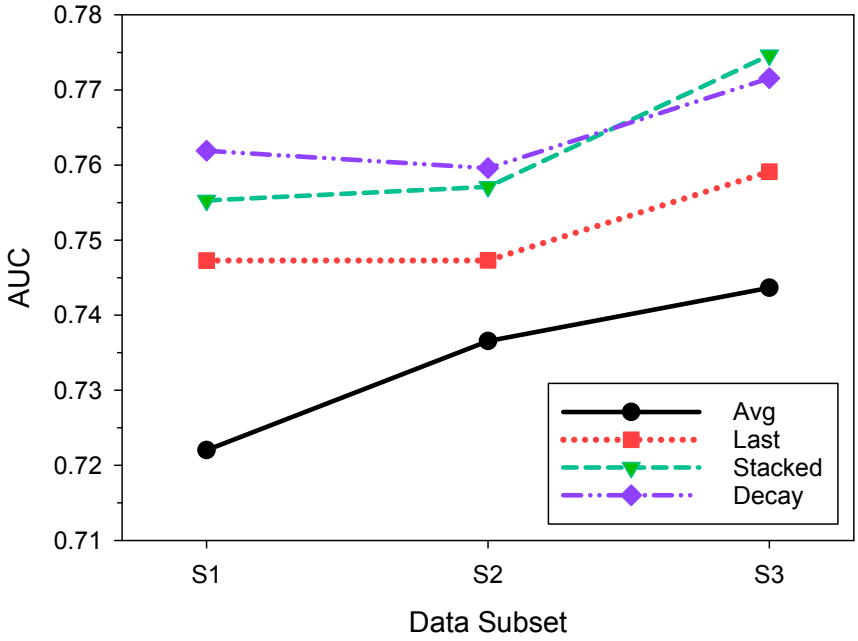

FIGURE 3.3: *Comparing representation extraction methods by classifiers in AUC (SVM-RBF).*

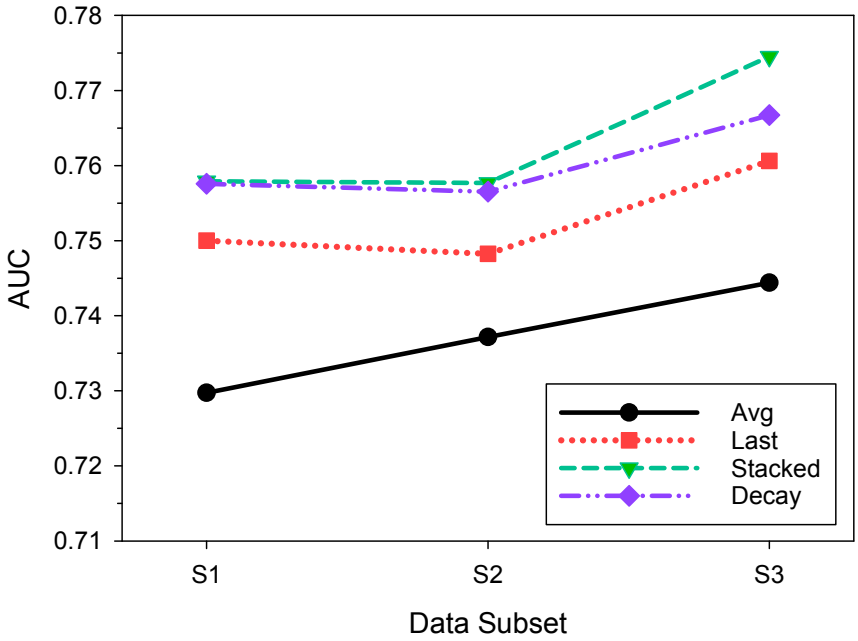

FIGURE 3.4: *Comparing representation extraction methods by classifiers in AUC (SVM-sigmoid).*

Representation Extraction Comparison

First, we compared four representation extraction strategies. The figures from Figure 3.2 to 3.4 show the results of SVM classifiers (with linear, RBF, and sigmoid kernels respectively) in AUC scores and

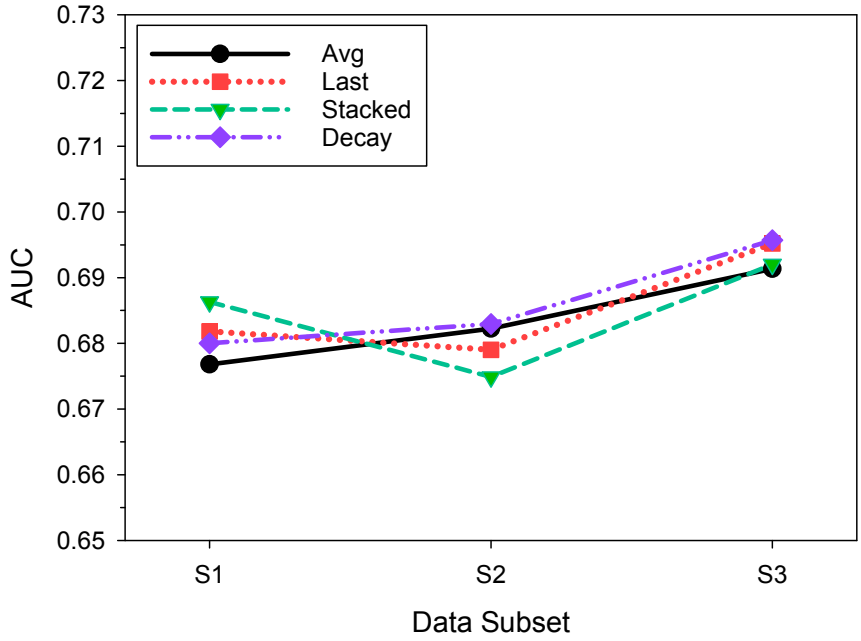

FIGURE 3.5: *Comparing representation extraction methods by classifiers in AUC (KNN).*

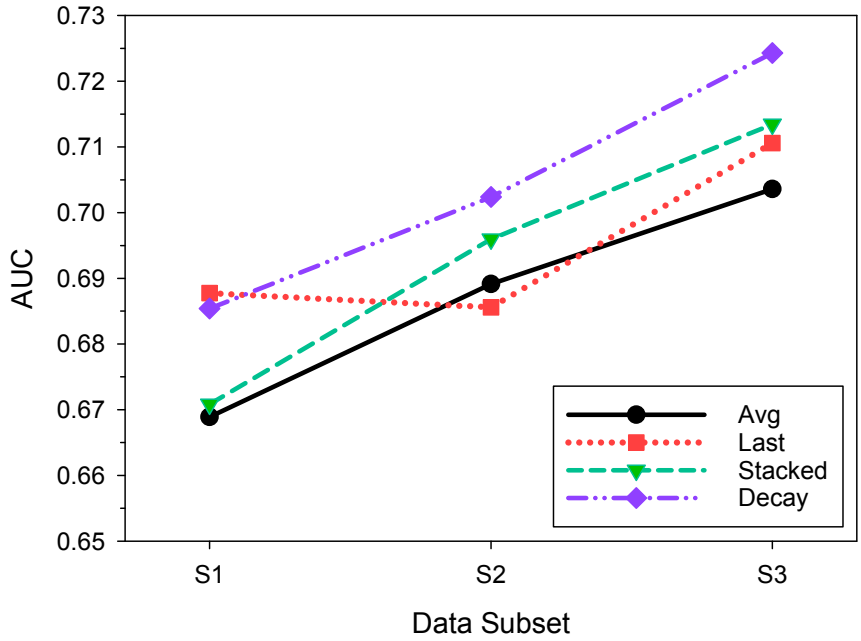

FIGURE 3.6: *Comparing representation extraction methods by classifiers in AUC (CART).*

the subsets are listed in a size-increasing order. Figure 3.5 and Figure 3.6 show the results for KNN and CART classifiers in AUC scores. The parameter α for *Decay* representation was set to 1. As a general trend, the performance of different methods increase with the increase of the subset size.

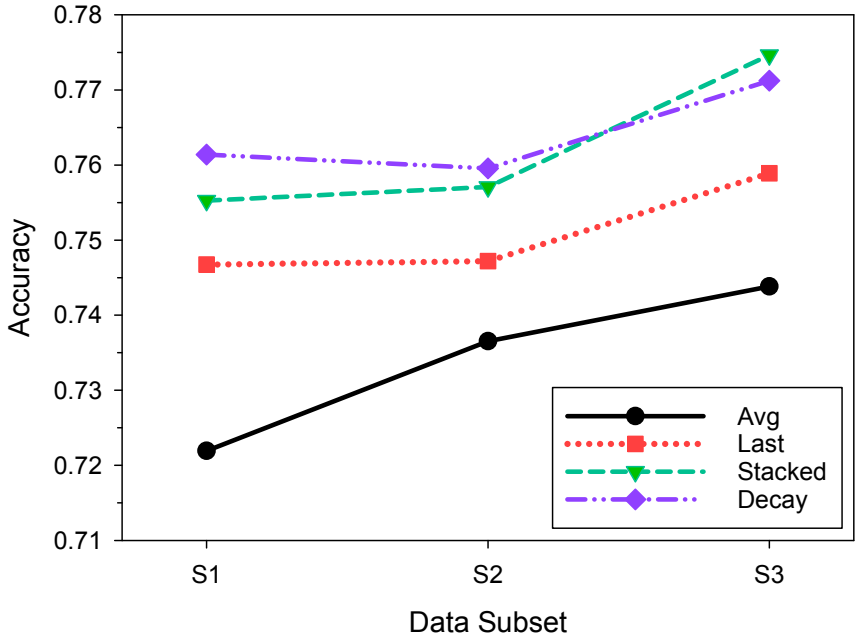

FIGURE 3.7: *Comparing representation extraction methods with SVM-RBF (Accuracy).*

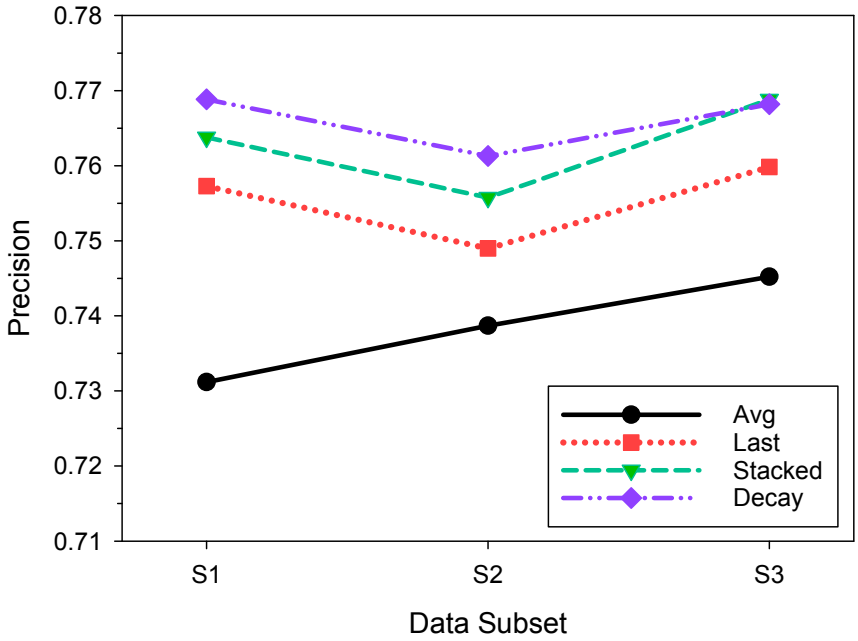

FIGURE 3.8: *Comparing representation extraction methods with SVM-RBF (Precision).*

The best performance was seen with *Stacked* representation at 77.46% AUC in Figure 3.3, achieved by SVM with RBF kernel on subset S_3. In general, the proposed *Stacked* and *Decay* representations constantly outperform the baseline *Avg* and *Last* representations on all subsets with SVM classifiers,

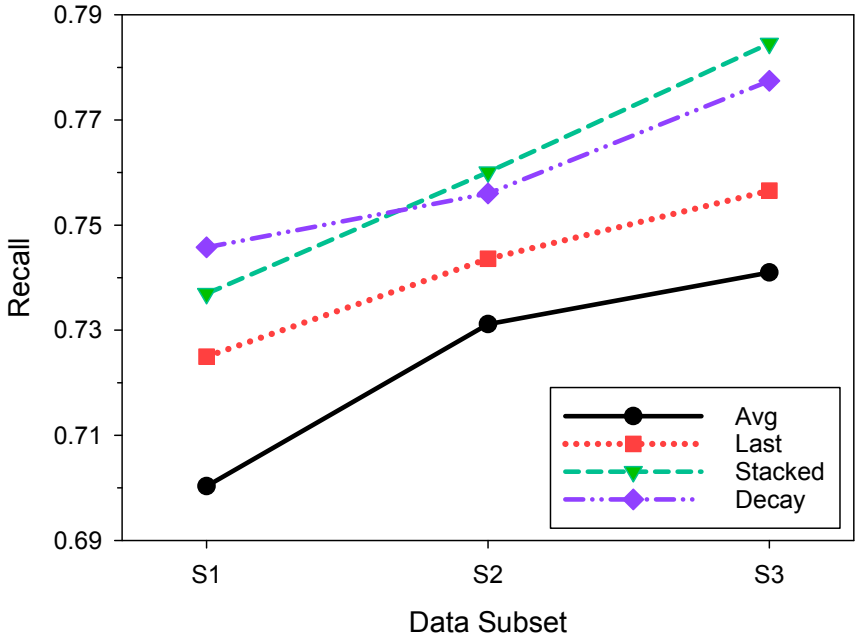

FIGURE 3.9: *Comparing representation extraction methods with SVM-RBF (Recall).*

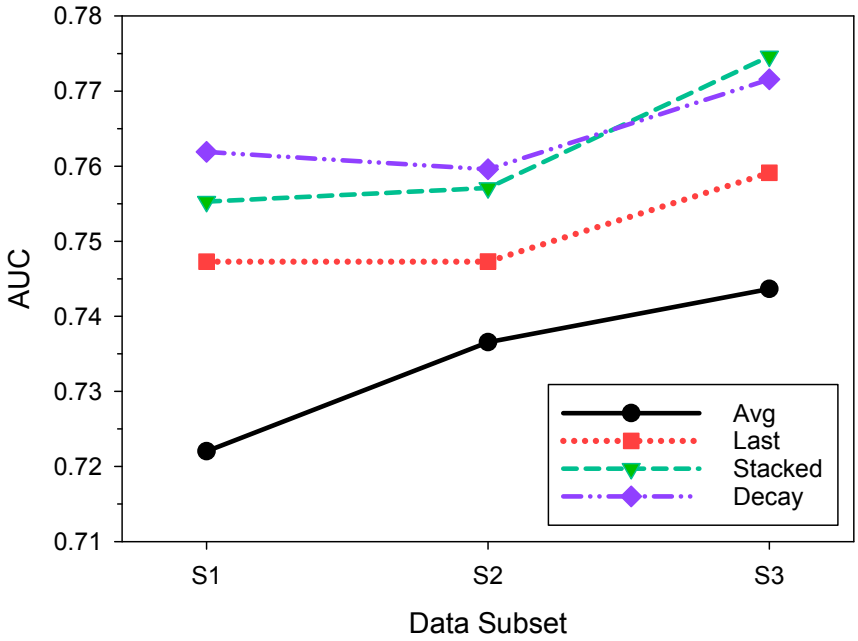

FIGURE 3.10: *Comparing representation extraction methods with SVM-RBF (AUC).*

although it was less obvious in the case of KNN and CART. Since SVM-RBF constantly gave the best performance, we will focus on its results in the following discussion.

The stability of the proposed representations can be further observed in Figure 3.7 to Figure 3.11,

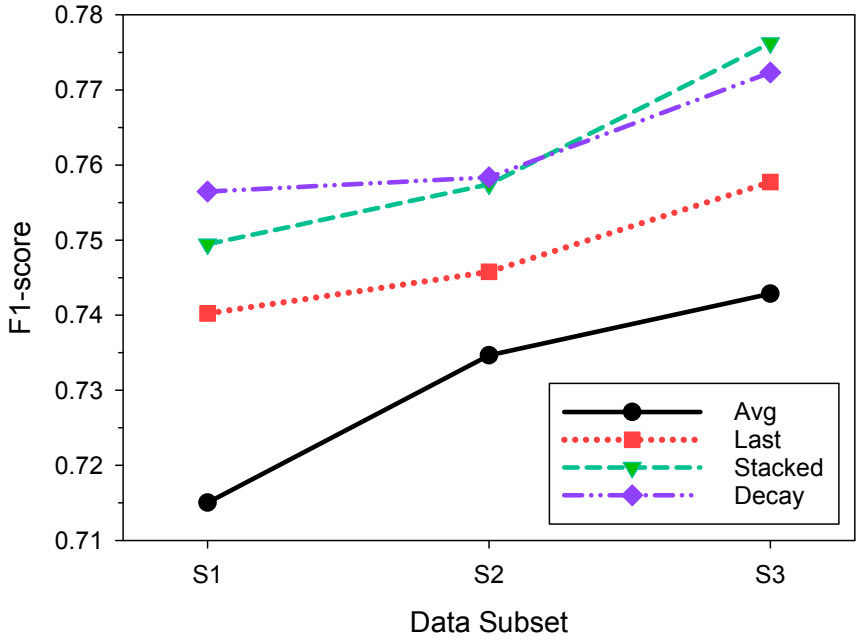

FIGURE 3.11: *Comparing representation extraction methods with SVM-RBF (F score).*

which show the performance in all evaluation measures. It is clear that the proposed *Stacked* and *Decay* consistently outperform baseline methods in all five measures and on all subsets. However, there is no clear winner between the *Stacked* and *Decay* methods.

The effect of setting parameter α for Eq. (3.4) are displayed in Figure 3.12 and Figure 3.13 in AUC scores. The range of values experimented were $\{1/25, 1/5, 1, 5, 25\}$. It can be observed that $\alpha = 1$ consistently delivers the best performance.

Feature Selection and Factor Analysis

To investigate the effects of feature selection and the selected subsets, we further conducted experiments with the Univariate F-score method (*F-univariate*) and the Logistic Regression model with L1 regularization (*LR-L1*) for feature selection. Table 3.2 to Table 3.4 show the selected feature sizes of *LR-L1* over C in the range $\{0.005, 0.01, 0.05, 0.1, 1\}$. The corresponding percentile (%) for feature selection is also given and *NoFS* denotes the base case without selection. Note that *Stacked* representation has twice as many features than *Decay*.

The figures from Figure 3.15 to Figure 3.17 show the results of feature selection on the proposed representations in AUC scores. The range of selection percentile for *F-univariate* is $\{5\%, 10\%, ..., 60\%\}$

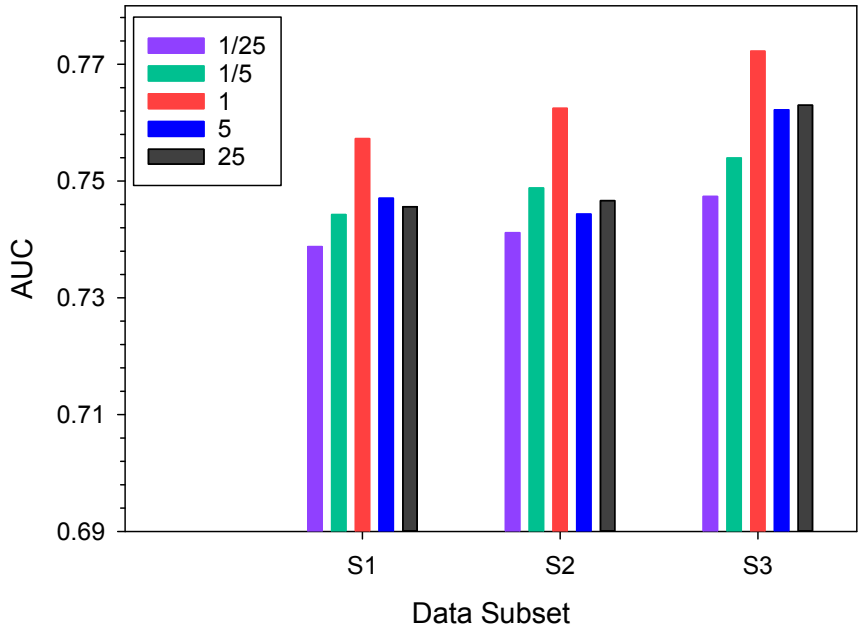

FIGURE 3.12: *Parameter α tuning for Decay feature representation in AUC, within the value range $\{1/25, 1/5, 1, 5, 25\}$ (LinSVM).*

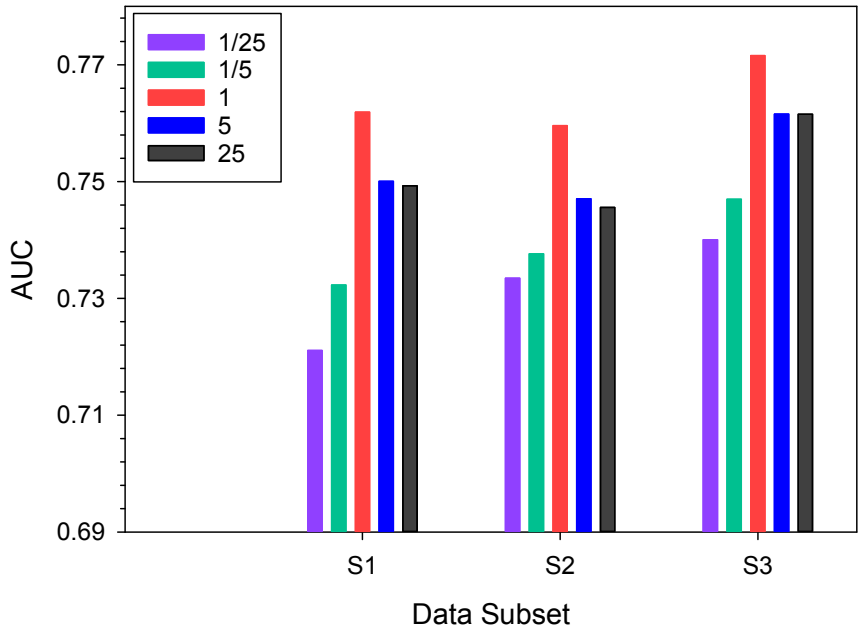

FIGURE 3.13: *Parameter α tuning for Decay feature representation in AUC, within the value range $\{1/25, 1/5, 1, 5, 25\}$ (SVM-RBF).*

and 100 percentile also refers to the base case without selection. It can be seen that *F-univariate* constantly improves the performance. The best performance was achieved at 77.98% on subset S_3,

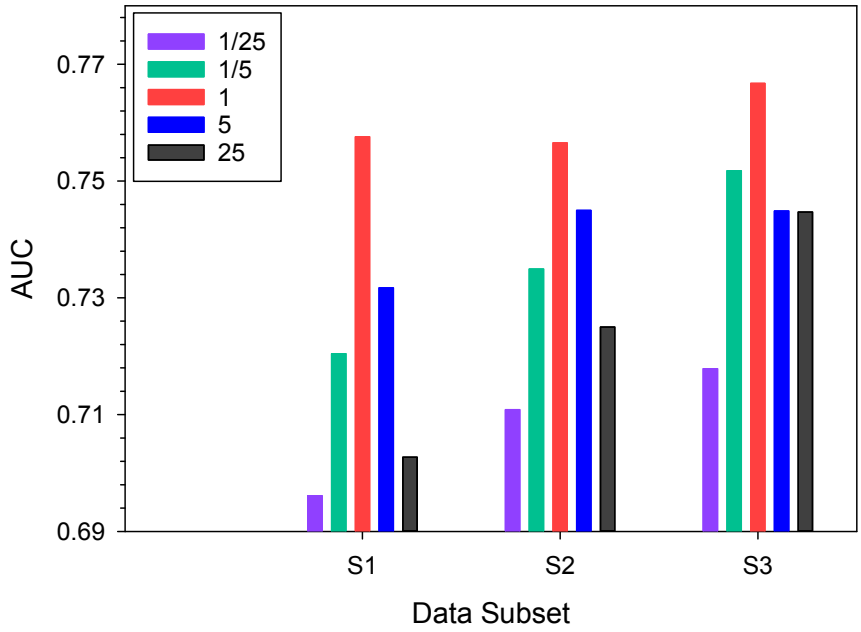

FIGURE 3.14: *Parameter α tuning for Decay feature representation in AUC, within the value range $\{1/25, 1/5, 1, 5, 25\}$ (SVM-sigmoid).*

TABLE 3.2: Average number of selected features by LR-L1 with a range of C values on Dataset S_1

	S_1			
C	Stacked	%	Decay	%
NoFS	178.0	100.0	89	100.0
1	106.9 ± 1.03	60.1	57 ± 2.06	64.1
0.1	30.5 ± 0.92	17.1	19.6 ± 0.85	22.1
0.05	18.7 ± 0.26	10.5	11.9 ± 0.57	13.4
0.01	3.7 ± 0.26	2.1	2.9 ± 0.32	3.3
0.005	2.6 ± 0.43	1.5	1.6 ± 0.53	1.7

TABLE 3.3: Average number of selected features by LR-L1 with a range of C values on Dataset S_2

	S_2			
C	Stacked	%	Decay	%
NoFS	178	100.0	89	100.0
1	125.3 ± 1.72	70.4	66.7 ± 2.28	37.5
0.1	45.8 ± 0.97	25.8	27.8 ± 1.32	15.7
0.05	22.1 ± 0.58	12.4	15.5 ± 0.52	8.7
0.01	7.0 ± 0.41	4.0	5.7 ± 0.52	3.2
0.005	3.0 ± 0.16	1.7	2.1 ± 0.01	1.2

0.51% higher than the base case, based on 50 percentile selection on the *Stacked* representation (Figure 3.17). Similarly, *Decay* obtains 77.87% on the same subset, and 0.75% improvement on the base case, based on only 20% of the features.

TABLE 3.4: Average number of selected features by LR-L1 with a range of C values on Dataset S_3.

	S_3			
C	Stacked	%	Decay	%
NoFS	178	100.0	89	100.0
1	127.4 ± 1.09	71.6	68.7 ± 2.01	77.2
0.1	46 ± 0.82	25.9	27.3 ± 1.16	30.7
0.05	28.4 ± 0.42	15.9	16.5 ± 0.97	18.5
0.01	9.2 ± 0.32	5.2	7.3 ± 0.43	8.1
0.005	3.1 ± 0.01	1.7	2.1 ± 0.01	2.3

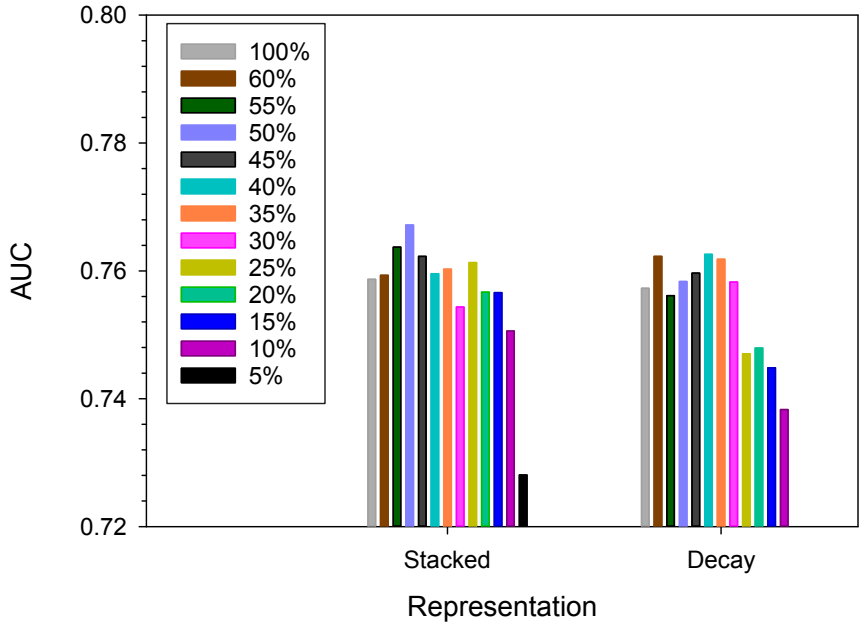

FIGURE 3.15: *Effects of F-univariate feature selection methods on subsets S_1 in AUC. The percentile of F-univariate is in the range of $\{5\%, 10\%, ..., 60\%, 100\%\}$, where 100 refers to no selection*

On the other hand, Figure 3.18 to Figure 3.20 show the results of feature selection with *LR-L1* on subsets $S_1 - S_3$ in AUC. The parameter C for *LR-L1* is in the range of $\{0.005, 0.01, 0.05, 0.1, 1\}$ and "NoFS" denotes no selection. It can be seen the *LR-L1* had no improvement on the performance; rather the performance drops with the decrease of the value of C.

We further investigated risk factors that played key roles in the classification. This information could provide explanations especially to those who were predicted with a low PHI. For the purpose of reporting selected features with a displayable size, we show the results of two selection methods with the selected size as close as possible in Table 3.5 and Table 3.6. Table 3.5 lists the nine features selected by *LR-L1* on the *Decay* representation with $C = 0.01$ and AUC 73.64%. Since the last

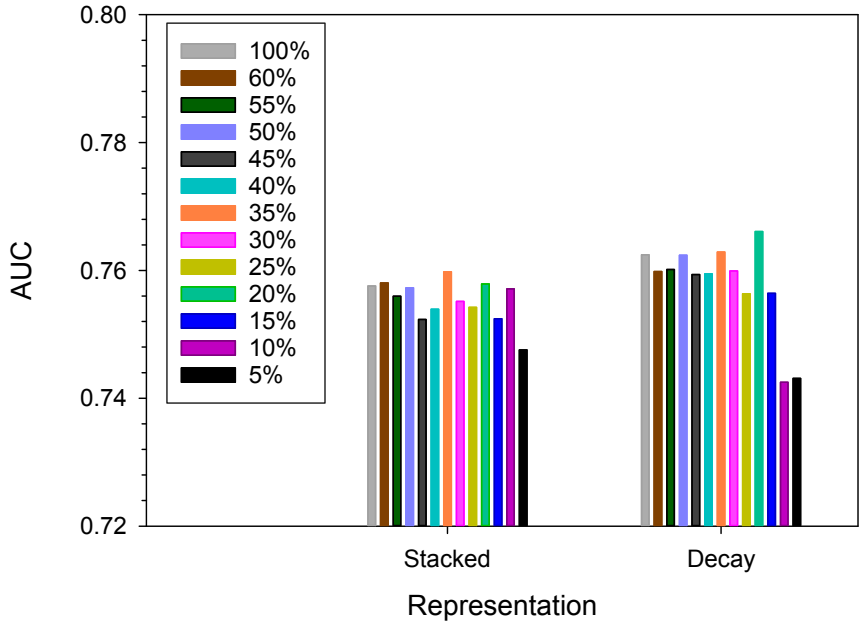

FIGURE 3.16: *Effects of F-univariate feature selection methods on subsets S_2 in AUC. The percentile of F-univariate is in the range of $\{5\%, 10\%, ..., 60\%, 100\%\}$, where 100 refers to no selection*

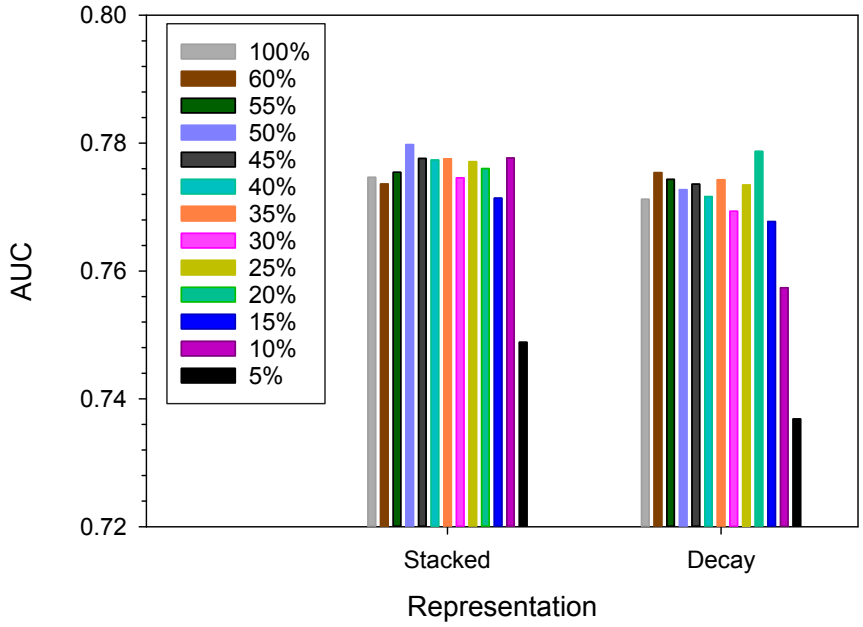

FIGURE 3.17: *Effects of F-univariate feature selection methods on subsets S_3 in AUC. The percentile of F-univariate is in the range of $\{5\%, 10\%, ..., 60\%, 100\%\}$, where 100 refers to no selection*

two features were only selected once out of ten in 10-fold cross validation, the averaged number of selected features is 8.1%. Table 3.6 shows the features selected by *F-univariate* at a slightly larger

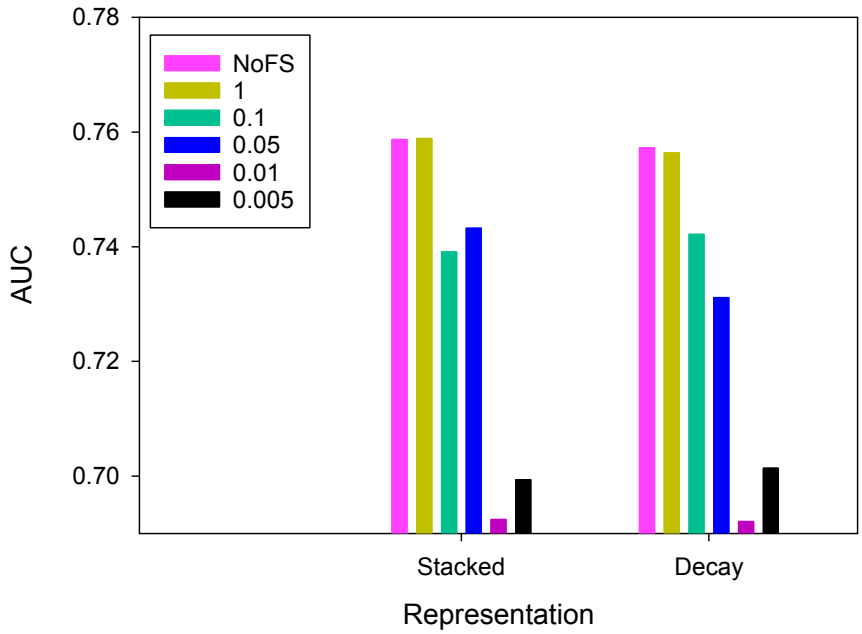

FIGURE 3.18: *Effects of LR-L1 feature selection methods on subsets S_1 in AUC. The parameter C for LR-L1 is in the range of $\{0.005, 0.01, 0.05, 0.1, 1\}$ and "NoFS" denotes no selection.*

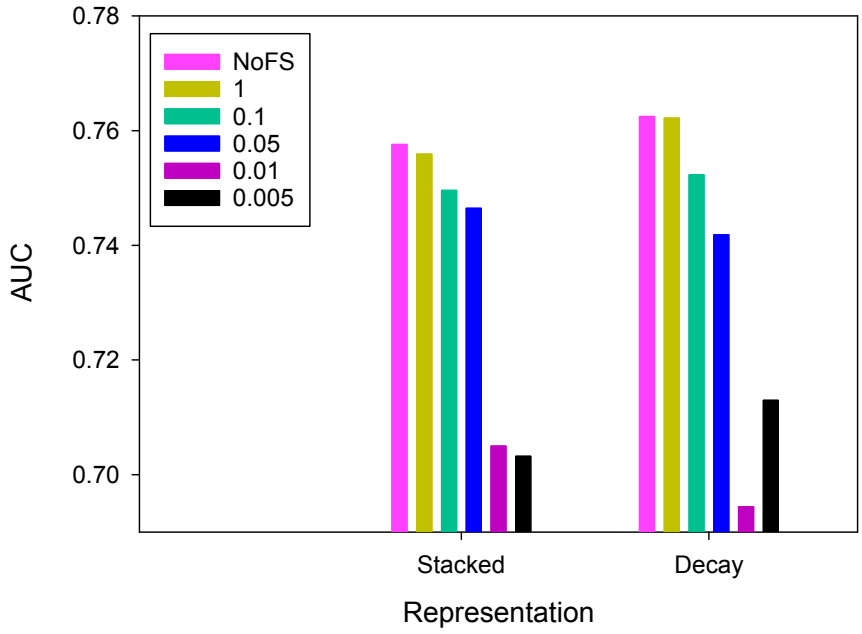

FIGURE 3.19: *Effects of LR-L1 feature selection methods on subsets S_2 in AUC. The parameter C for LR-L1 is in the range of $\{0.005, 0.01, 0.05, 0.1, 1\}$ and "NoFS" denotes no selection.*

percentile, i.e., 10%, with an AUC score of 75.34%. The lists are ordered by feature scores output by selection methods.

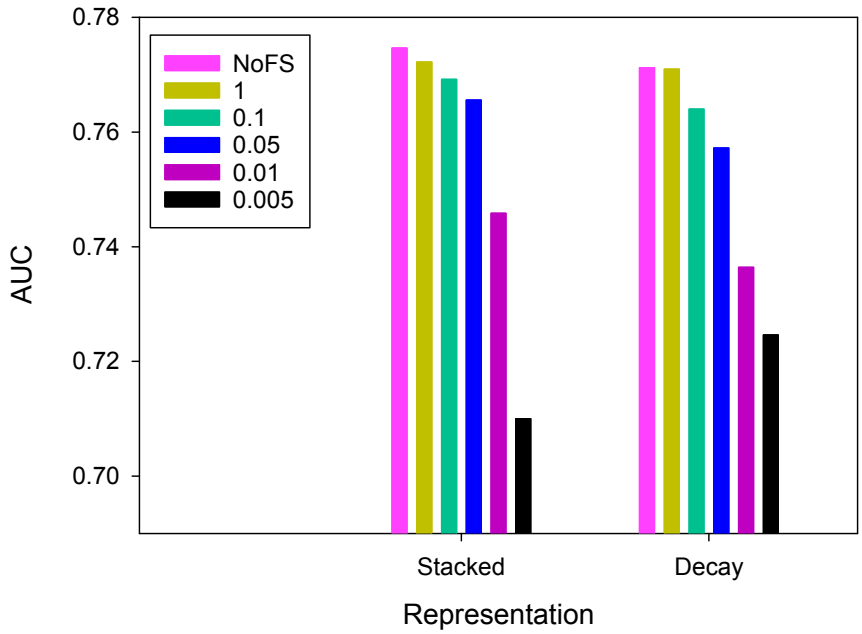

FIGURE 3.20: *Effects of LR-L1 feature selection methods on subsets S_3 in AUC. The parameter C for LR-L1 is in the range of $\{0.005, 0.01, 0.05, 0.1, 1\}$ and "NoFS" denotes no selection.*

TABLE 3.5: Selected risk factors by $LR - L1$ feature selection method.

	LR-L1	
	Factor	score (AbsAvg $\pm$ std)
1	**hemoglobin (hb)**	0.73 ± 0.0651
2	**age**	0.38 ± 0.0061
3	**SPMSQ**	0.37 ± 0.0234
4	**tsh**	0.16 ± 0.0906
5	exercise	0.09 ± 0.0157
6	BMI	0.07 ± 0.0185
7	education level	0.02 ± 0.0043
8	**hematocrit (hct)**	$< 0.001 \pm 0.0013$
9	**mch**	$< 0.001 \pm 0.0005$

Out of the nine features selected by these two methods, six of them overlapped, i.e., *hemoglobin (hb), age, SPMSQ cognitive function assessment, thyroid stimulating hormone (tsh), hematocrit (hct), and mean corpuscular hemoglobin (mch).* Amongst them, three were ranked top four by both methods. *Hb* test, for example, reveals the blood glucose levels, which is an important indicator for diabetes. Although *age* is a known factor for mortality, Table 3.5 and Table 3.6 show that its combination with other factors, say poor cognitive functions (i.e., SPMSQ), led to high mortality risks.

TABLE 3.6: Selected risk factors by *F-univariate* feature selection method.

	F-univariate	
	Factor	score (AbsAvg $\pm$ std)
1	**age**	781 $\pm$ 17.18
2	**hemoglobin (hb)**	769 $\pm$ 24.39
3	**hematocrit (hct)**	669 $\pm$ 21.21
4	**tsh**	574 $\pm$ 23.57
5	**SPMSQ**	351 $\pm$ 6.95
6	**mch**	342 $\pm$ 11.20
7	albumin (alb)	307 $\pm$ 7.12
8	waist	260 $\pm$ 9.87
9	red blood cell	239 $\pm$ 18.89

Personal Health Index Applications

To illustrate the usefulness of PHI, we first include a dashboard of regional health in Figure 3.21. It shows the averaged PHIs of the 12 districts of Taipei City computed based on the full GHE dataset. On the right-hand map, the corresponding regions are colored based on a color spectrum from red to green denoting PHI in the [0.5,1.0] range. The districts are indexed by their zip codes. The average PHI is 0.8852$\pm$0.008 with standard deviation. Such information can reveal regional health conditions and assist local government's fine-tuning health policies.

In addition, Figure 3.22 presents a dashboard of a person's health status summary, where yearly PHI scores are displayed on the top right chart. Detailed examination results can be compared across years via the PHI fingerprint on the left as a concentric circle. It shows the severity of individual test results, where an annulus represents the results of a year.

3.3.5 Discussion

From the results demonstrated in Section 3.3.4, we have demonstrated the effectiveness of our proposed classification-based PHI prediction framework. We compared the performance of SVM classifiers with linear, RBF, and sigmoid kernels, as well as KNN and CART classifiers. The best prediction was achieved at 77.98% AUC on the proposed *Stacked* representation plus *F-univariate* (Figure 3.15 to Figure 3.17). Although our experiments were based on one dataset, it is a large and comprehensive GHE dataset with 102,258 participants.

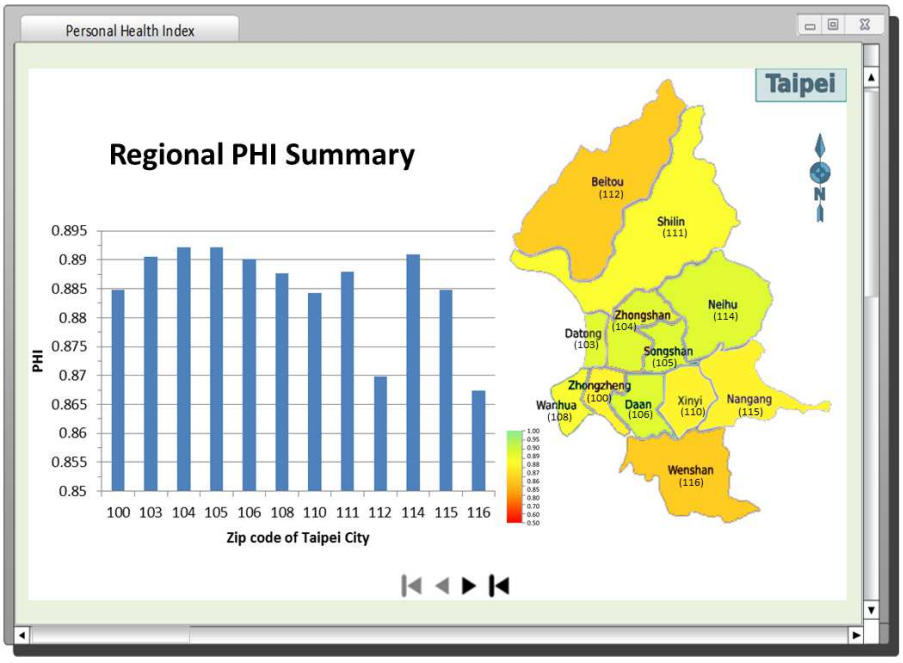

FIGURE 3.21: *A dashboard of Taipei City's regional health based on PHI. The left-hand chart displays the regional PHI of twelve districts of Taipei City. The map on the right has the corresponding districts coloured based on a colour spectrum from red to green denoting PHI in the [0.5,1.0] range. The districts are indexed by their zip codes.*

The proposed *Stacked* and *Decay* representation extraction strategies were also shown to consistently outperform the baseline *Avg* and *Last* methods in all five measures (Figure 3.2 to Figure 3.11). This verifies the intuition that the latest GHE record is most indicative while historical records are still accountable for health status prediction.

The similar performances of *Stacked* and *Decay* suggest that weighting historical records equally or as a time-decay function has little effect on the performance. Also from the fact that *Stacked* has twice as many dimensions than *Decay*, we conclude that *Decay* is preferred when dimensionality of given data is high.

F-univariate feature selection was shown to constantly improve performances (Figure 3.15 to Figure 3.17), while *LR-L1* dragged the performances down (Figure 3.18 to Figure 3.20). We further showed displayable lists of selected features in Table 3.5 and Table 3.6. This information can be useful for providing explanations especially to those predicted with low PHI. Finally, two application examples of PHI were given in Figure 3.21 and Figure 3.22, which show the applicability of PHI at both population and individual levels.

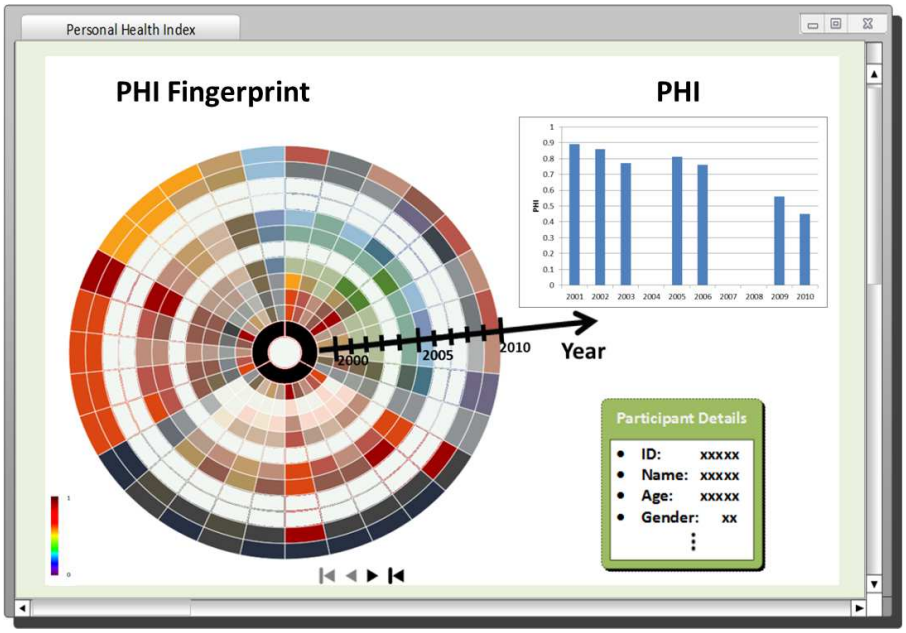

FIGURE 3.22: *A dashboard of a person's health status summary. PHI is computed based on GHE results for each year and summarized in the bar chart on the right. PHI fingerprint on the left is a breakdown of PHIs into details, where an annulus of the concentric circle represents the results of a year and a sector in an annulus denotes a test result of the year. Colour indicates severity and the white annuluses denote no attendance.*

3.4 Summary

Computing a comprehensive health score for every citizen was impossible before the *big data* era. Because of the wide availability of multiple sources of data collected from all kinds of devices over many years, effectively evaluating health status of a person from-cradle-to-grave is becoming possible. One example of such a data source is the annual Geriatric Health Examination (GHE) which is now an integral part of elderly healthcare for many developed countries.

Predicting personal health status based on health examinations reveals a promising and important trend in healthcare research. In this work, we took the first step by making such predictions from longitudinal GHE records of 102,258 participants. Particularly, we defined the Personal Health Index (PHI) to quantize health status and developed a classification-based prediction framework to effectively infer a person's PHI.

The effectiveness and stability of the proposed framework have been demonstrated via our extensive experiments. The best prediction was achieved at 77.98% AUC on the proposed *Stacked* representation method with feature selection. It has also demonstrated that the proposed methods

consistently outperform the baselines in all five evaluation measures.

Chapter 4

Disease-level Classification with Label Uncertainty

4.1 Problems and Challenges

In the last decade, an increasing number of data mining applications have been developed to support healthcare decision making [37, 134]. Of particular focus in recent years has been studies in the area of clinical risk classification [122, 121, 127, 88, 92]. These studies generally treat class labels with 100% certainty. However, label uncertainty is commonly found in clinical judgments due to subjectivity and inadequate information [104]. Often it has been handled as noise, and the task has been to detect and correct mislabeling [87, 107, 45]. However, in the case of multiple, non-exclusive medical conditions [82], such as comorbidity, it makes more sense to treat labels with degrees of certainty rather than forcing them to belong to one "true" class.

Recently in the field of Computer Vision, Yi *et al.* introduced a soft-label learning model for complex event detection on Web videos [133]. For a given small number of target event instances, the model leverages instances where their relationship to the target event is uncertain and to be learned. Inspired by this work, we formulated the PHI prediction problem as a soft-label optimization problem.

In the process of soft-label learning, we distinguished the following three types of instances:

- participants with a target COD label,

- participants with a non-target COD label, and
- participants without a COD label.

Note that in the above list, a "target" COD refers to the COD of interest to be predicted.

Traditionally, in a binary prediction problem on a target event, instances of the first type are regarded as positive and those of the other two types, namely those with a non-target labels and those without a label, are treated as negative. However, in our case, participant records with different COD labels might share similar traits due to comorbidity, so some non-target instances could be "related" to the target ones. The method we propose is designed to capture the differences amongst these three types of instances.

This work substantially improved our previous work presented in Chapter 3.

- Firstly, we extended the concept of PHI from a single overall health score based on all-cause mortality to a vector of scores, each reflecting personal health risk in a disease category; and
- Secondly, rather than treating labels as 100% certain, as in our previous work, we took a soft-label learning approach to handle label uncertainty.

Experimentally we demonstrated the effectiveness of MyPHI based on a large Geriatric Health Examination (GHE) data set of 262,424 records from 102,258 participants.

The rest of the chapter is organized as follows. Section 4.2 details our MyPHI method and highlights the optimization technique employed to construct the prediction model. In Section 4.3, we demonstrate the effectiveness of our method through extensive experiments. Section 4.4 is our summary and findings.

4.2 The Methodology

The proposed MyPHI prediction method that computes Personal Health Index (PHI) for elderly healthcare contains three key components, namely data pre-processing, feature extraction, and model learning with uncertainty, as shown in Figure 4.1. The inputs of the method are the Geriatric Health Examination (GHE) records of a population linked to the main Cause of Death (COD) database. The

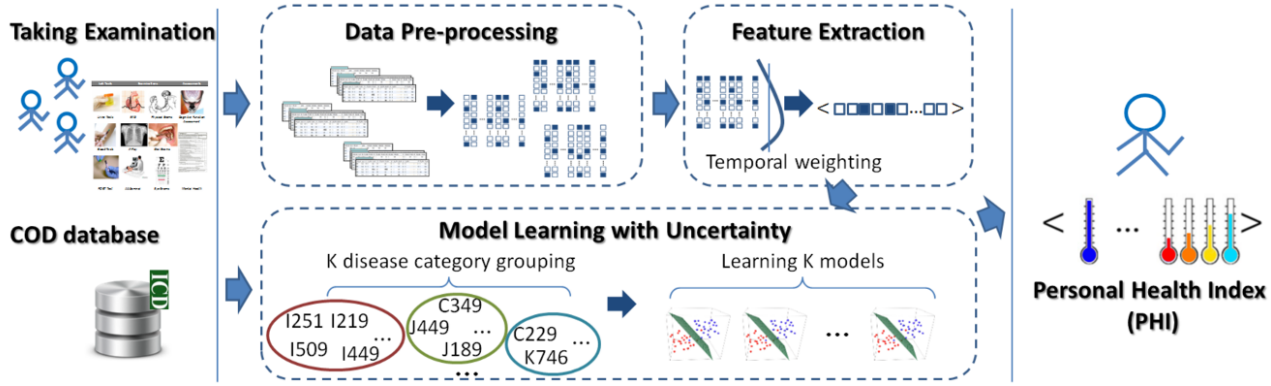

FIGURE 4.1: The process of Personal Health Index (PHI) prediction

output is a vector of k predicted scores in the range of $[0, 1]$ interval, reflecting personal health risks in k disease categories. Note that the algorithm is designed for medical datasets that share the characteristics described in Section 1.2.2 of Chapter 1. Although in the following discussions we will use the GHE dataset as an example, the applicability of our proposed method is not limited to that dataset.

4.2.1 Data Pre-processing

The datasets used in this work include a Geriatric Health Examination (GHE) dataset linked to a Cause of Death (COD) dataset. For detailed descriptions of these datasets and their characteristics, please refer to Section 1.2.1.

In this work, we adopted an event-based view that treated an observed abnormal result as an occurrence of an abnormal event. This approach was based on the practice of evidence-based medicine that only took the observed symptoms and signs into account in clinical diagnosis. For binary variable, abnormality was encoded as 1, and 0 otherwise. Real values were firstly discretized into bins. Ordinal and categorical variables were binarized into a vector of binary variables representing the unique values of the original variables. A variable took the value 1 if the original variable took the corresponding value; otherwise, it was 0. For the details of data preprocessing, especially about handling attributes of different data formats, please refer to Section 3.2.1 of Chapter 3. Since GHE records are essentially longitudinal, participants with only one record are excluded from consideration. This left us with a subset of 221,074 records from 60,881 participants.

After the pre-processing stage, participant longitudinal records could be represented as a sequence of time-stamped records, which then can be formally defined as follows:

Definition 4.1 (Record sequence)**.** *A record sequence s_i of a person p_i is an ordered list of m records $\langle r_{i1}, ..., r_{ij}, ..., r_{im} \rangle$, where record r_{ij} is a tuple (t_{ij}, v_{ij}) of a d dimensional binary vector v_{ij} of values observed at time t_{ij}, where t_{ij} is the normalized time mapped onto an integer space such that $t_{ij} \in \mathbb{N}$ and $t_{ij} < t_{ij+1}$.*

4.2.2 Feature Extraction

One of the key challenges we faced was how to model and represent a participant's record sequence in our framework. To begin with, a participant could have multiple records as depicted earlier in Figure 1.1 of Chapter 1, so it is not naturally in the form of a feature vector. However, these records cannot be simply flattened into a feature vector because record sequences of participants have varied lengths. On the other hand, these record sequences were not like the traditional time series such as ECG series, because they were infrequent, as discussed in Section 1.2.2. In fact, the average number of records per participant was 2.56.

In this work, we took a feature-based approach [129] that converted sequences into a point-based representation, i.e., by transforming a sequence into a vector of features. So the problem became, "How do we design a transformation mechanism that has a greater ability to capture different shapes of curves?"

Our previous work on representation extraction strategies (Chapter 3, [22]) suggested that time smoothing kernels that assign time weight to values at time t outperformed methods without considering the longitudinal progression. Based on the previous results, we designed a chi-squared kernel in our current work to model the changes of importance over time. This was chosen over the commonly used Gaussian distribution because it had a greater ability in capturing different shapes of curves.

The probability density function of the chi-squared distribution is defined as:

$$f(t, \theta) = \begin{cases} \frac{t^{\theta/2-1} e^{-t/2}}{2^{\theta/2}\Gamma(\theta/2)} & t \geq 0 \\ 0 & \textit{otherwise} \end{cases} \tag{4.1}$$

Algorithm 1 Feature Extraction

Input: S: a list of record sequences, θ: a scale parameter for the temporal weighting kernel, T: the time window of interest.
Output: X: extracted feature vector of S.
$X := []$
while $i < size(S)$ **do**
$((t_{ij}, v_{ij}))_{j \in 1 \ldots n_i} := S(i)$
$X(i) := \sum_{j=1}^{n_i} K_\theta(T - t_{ij} + 1) \cdot v_{ij}$
end while
return X

where θ is the degree of freedom, $\Gamma(\cdot)$ is the Gamma function, and t in our case is a parameter that takes integers greater than zero.

The chi-squared kernel is defined as a function of truncated chi-squared distribution:

$$K_\theta(t) = \begin{cases} \frac{f(t,\theta)}{\Phi(T)-\Phi(1)} & t \in [1, T] \\ 0 & \textit{otherwise} \end{cases} \tag{4.2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $f(\cdot)$ and T the window length.

The record sequence s_i of a person p_i is then transformed into a vector x_i using Eq.(4.2):

$$x_i = \sum_{j=1}^{m} K_\theta(T - t_{ij} + 1) \cdot v_{ij} \tag{4.3}$$

where the $(T - t_{ij} + 1)$ term reverses the time ordering, resulting in a larger weight to the latest records; v_{ij} is the j^{th} record of s_i obtained at time t_{ij}.

The full procedure of feature extraction is summarized in Algorithm 1. It takes a list of record sequences S of length n, a scale parameter θ for the temporal weighting kernel, and T as the time window width of interest as inputs. It iterates through the sequence list and transforms every sequence into a feature vector using the weighting kernel based on the scale parameter. The output is a $n \times m$ matrix X, where m is the dimension of the feature vector.

4.2.3 Model Learning with Label Uncertainty

To build an effective PHI prediction model for the elderly, we trained k models, each for a target disease category. As discussed earlier in Section 4.1, there can be a degree of uncertainty that a COD label is assigned to a particular person. This uncertainty can be observed in two types of instances, namely the target disease instances (known as the positive examples), and the non-target disease instances (usually treated as the negative examples). Intuitively, different people can belong to a target disease in different degrees. In addition, non-target cases and alive cases, though all considered negative examples in the traditional sense, are "negative" in different ways: non-target cases may be closer to the target cases in the feature space than to the alive cases.

Optimization Problem

Due to the uncertainty of labels, we formulated the problem as an optimization problem [133] that finds a soft label for every instance. More specifically, given training instances $X = \{x_1, ..., x_n\}$ converted from n record sequences using Eq. (4.3), the soft label Y_i of x_i is $1 + S_i$ for target instances, $1 - S_i$ for non-target instances, and 0 for alive instances, where S_i expresses a degree of certainty to be learned.

This design can be expressed by defining:

$$Y = Y^a + A \odot S$$

where $A \odot S$ is the entrywise product of vectors $A \in \mathbb{R}^n$, $S \in \mathbb{R}^n$, $S \geq 0$, and Y_i^a is defined as follows:

$$Y_i^a = \begin{cases} 1 & \text{for the target and non-target instances} \\ 0 & \text{otherwise} \end{cases}$$

and A_i is defined as follows:

$$A_i = \begin{cases} 1 & \text{for the target instances} \\ -1 & \text{for the non-target instances} \\ 0 & \text{otherwise} \end{cases}$$

So the optimization problem can be defined as a regularized least squared minimization problem with additional constraints on Y and S:

$$\begin{aligned} &\min_{P,S,Y} ||X^{\mathrm{T}}P - Y||_F^2 + \Omega(P) \\ &s.t.\ Y = Y^a + A \odot S,\ S \geq 0 \end{aligned} \tag{4.4}$$

where $\Omega(\cdot)$ is a regularization term on P to prevent over-fitting and $||\cdot||_F$ denotes the Frobenius norm over a $m \times n$ matrix B such that:

$$||B||_F = (\sum_{i=1}^{m}\sum_{j=1}^{n} |b_{ij}|^2)^{1/2}$$

We further postulate that noises of the labels were mostly from the non-target disease cases, as these cases could exhibit similar or completely different traits to the target cases. So a weight vector W was introduced, which is learned only using the target disease instances and alive instances. By applying trace norm $||\cdot||_*$ on $E = [W, P]$, P can be further constrained by W, as the trace-norm can uncover the shared knowledge of W and P [132]. The optimization problem is modified as:

$$\begin{aligned} &\min_{W,P,S,Y} ||\tilde{X}^{\mathrm{T}}W - \tilde{Y}||_F^2 + ||X^{\mathrm{T}}P - Y||_F^2 \\ &\quad +\alpha(||W||_F^2 + ||P||_F^2) + \beta||E||_* \\ &s.t.\ Y = Y^a + A \odot S,\ E = [W, P],\ S \geq 0 \end{aligned} \tag{4.5}$$

where $\tilde{X}$ is the input data for the target and alive cases only, W and P are regularized using the Frobenius norm to prevent over-fitting, and together (i.e., E) they are further regularized using the trace-norm. α and β are the coefficients for the regularization terms, and $\tilde{Y}$ is the corresponding

label:

$$\tilde{Y}_i = \begin{cases} 1 & \text{for the target cases} \\ 0 & \text{for the alive cases} \end{cases}$$

Optimization Procedure

We now describe the procedure of solving the optimization problem formulated in Eq. (4.5). Let

$$D = \tfrac{1}{2}(EE^{\mathrm{T}})^{-\frac{1}{2}}$$

Eq. (4.5) can be converted to:

$$\begin{aligned} &\min_{W,P,S,Y} ||\tilde{X}^{\mathrm{T}}W - \tilde{Y}||_F^2 + ||X^{\mathrm{T}}P - Y||_F^2 \\ &+\alpha(||W||_F^2 + ||P||_F^2) + \beta Tr(E^{\mathrm{T}}DE) \\ &s.t.\ Y = Y^a + A \odot S,\ E = [W, P],\ S \geq 0 \end{aligned} \tag{4.6}$$

Eq. (4.6) can be solved through iteratively updating P, W, and S until convergence, by setting their partial derivatives to zero, one at a time, and solving it accordingly. The convergence has been proven in [133]. By setting the derivative of Eq. (4.6) w.r.t. P to 0, we have:

$$P = (XX^{\mathrm{T}} + \alpha I + \beta D)^{-1}XY \tag{4.7}$$

Again, by fixing P and setting the derivative of Eq. (4.6) w.r.t. W to 0, we get:

$$W = (\tilde{X}\tilde{X}^{\mathrm{T}} + \alpha I + \beta D)^{-1}\tilde{X}\tilde{Y} \tag{4.8}$$

Optimizing S is to solve the following problem:

$$\min_{S \geq 0} ||X^{\mathrm{T}}P - (Y^a + A \odot S)||_F^2 \tag{4.9}$$

Algorithm 2 Learning with Uncertain Labels

Input: $X \in \mathbb{R}^{d \times n}$: extracted d dimensional feature vectors of n participants, $\tilde{X} \subseteq X$: a subset of X of size m, containing only the target and alive cases, $\tilde{Y} \in \mathbb{R}^{m \times 1}$: the corresponding labels of $\tilde{X}$, $Y^a \in \mathbb{R}^{n \times 1}$ and $A \in \mathbb{R}^{n \times 1}$: parameters for learning uncertain labels.

Output: Optimized W, P, S.

Set $t = 0$ and initialize W, P randomly;

repeat

Compute D_t as: $D_t = \frac{1}{2}(E_t E_t^{\mathrm{T}})^{-\frac{1}{2}}$;

Update P_t according to Eq. (4.7);

Update W_t according to Eq. (4.8);

Compute $M_t = X^{\mathrm{T}} P_t - Y^a$;

Compute S_t by $S_{tij} = max(M_{tij}/A_{ij}, 0)$ (Eq. (4.11));

$t = t + 1$

until *Convergence*

Let $M = X^{\mathrm{T}} P - Y^a$. The problem then becomes:

$$\min_{S \geq 0} ||M - A \odot S)||_F^2 \tag{4.10}$$

Finally, the optimal solution to Eq. (4.10) is obtained by:

$$S_{ij} = \max(M_{ij}/A_{ij}, 0) \tag{4.11}$$

As shown in Algorithm 2, the optimal solution to Eq. (4.6) is obtained by iteratively updating P, W, S with Eq. (4.8) - Eq. (4.10) until convergence.

Given an example x_t, the predicted score for a target disease is $P^{\mathrm{T}} x_t$. Let all the k models (i.e., each is a P) be stored in $\Lambda \in \mathbb{R}^{d \times k}$. The prediction function can then be defined as:

$$f(x_t) = \Lambda^{\mathrm{T}} x_t \tag{4.12}$$

where $f : \mathbb{R}^{d \times 1} \rightarrow [-1, 1]^{k \times 1}$.

4.2.4 Personal Health Index Calibration

To convert the predicted scores output by Algorithm 2 into probabilities and allow a person's PHI to be comparable to that of others, we further employ a step of PHI calibration.

First, the scores are z-normalized within the model outputs:

$$z_k = \frac{f_k - \mu_k}{\sigma_k} \tag{4.13}$$

where f_k is the k^{th} score of the output vector, and μ_k and σ_k are the mean and standard deviation of the k^{th} model outputs.

Since those with high risks are in the extreme end of the spectrum, we employ a generalized extreme value distribution [53]:

$$G(t) = e^{-[1-t]}$$

where t indicates the time. Note that it has a steeper growth when $t > 0$. So the final PHI calibration function for the k^{th} model outputs is:

$$G_k(z_k) = e^{-[1-z_k]} \tag{4.14}$$

Finally, the k^{th} score of PHI is the compliment probability:

$$PHI = 1 - G_k(z_k) \tag{4.15}$$

4.3 Experiments and Results

In this section, we report and analyze our experiment results. Extensive experiments were conducted to evaluate MyPHI using a real-world GHE dataset described in Section 1.2.1. We first discuss how the 10 disease classes were selected, followed by the experiment setup. Then we analyse and discuss the results.

4.3.1 Disease Category Grouping

We selected top 10 disease categories that have the highest frequency counts in the GHE dataset based on the linked Cause of Death (COD) labels encoded in ICD9 and ICD10. Table 4.1 shows the list

TABLE 4.1: Numbers of positive cases in disease categories

Top k	Disease Category	Count
1	Lung	649
2	Heart	296
3	Cerebrovascular	153
4	Diabetes	112
5	Stomach	105
6	Colon	101
7	Liver	83
8	Pancreas	61
9	Septicaemia	60
10	Hypertension	42
11	Other	1,314

of those top 10 disease categories. We designed an 11th class called the "Other" category, which contains all the other health-related ICD codes not in the top 10 disease categories. The sizes of the 11 classes are also shown in Table 4.1.

4.3.2 Experiment Setup

We compared MyPHI with two typical classification methods commonly used in medical applications as baselines, namely the Support Vector Machines model with linear kernel (LinSVM) and the Logistic Regression (LR) model. We used LIBSVM [18] for the implementation of LinSVM and LIBLINEAR [39] for the implementation of LR.

In addition, we compared MyPHI with the class-weighted versions of the baseline methods, denoted as LinSVM-W and LR-W respectively, where the class weights were set according to the ratio of positive and negative class.

We trained a model for each of the 11 disease categories, following the 35:35:30 stratified train/validate/test split ratio in all experiments. To investigate the effect of class imbalance, the negative (alive) cases were sub-sampled according to the positive vs. negative ratios 1:1, 1:10 and 1:100. For the "Other" disease category in the case of 1:100, since the portion of negative size exceeds the total number of negative cases, we report only the results of 10 disease categories.

The parameters of all the algorithms were searched on the grid of $\{10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}\}$ using

TABLE 4.2: The averaged AUC (%) of 11 disease categories of various positive vs. negative ratios. The proposed MyPHI significantly outperforms all the other algorithms in most cases.

	Clean Case					Noisy Case				
Ratio	MyPHI	LinSVM	LinSVM-W	LR	LR-W	MyPHI	LinSVM	LinSVM-W	LR	LR-W
1:1	**83.48**	82.11	82.11	82.79	82.79	**74.07**	71.23	71.23	70.47	70.47
1:10	85.65	77.99	**85.66**	78.72	85.19	**85.14**	75.03	81.56	75.65	80.32
1:100	**89.95**	68.57	85.91	70.42	85.99	**89.37**	69.1	83.83	70.45	84.04

the validation set. The parameter θ for chi squared kernel in feature extraction phase was experimentally set to 4.

4.3.3 Results

In this subsection, we analyse our experiment results. We first compare the overall results of all algorithms in two settings, namely a clean case and a noisy case. Secondly, we look into the performance of the prediction on individual disease categories. Then, we investigate the effects of having the trace-norm in our algorithm. Finally, we show the effects of the length of record sequence on the performance.

Clean vs. Noisy Cases

The algorithms were firstly evaluated under the ideal situation, where there were only positive instances (i.e., those whose main cause of death is the target disease) and negative instances (i.e., those who are alive). We call this setting the clean case, as there are no non-target instances to confuse the learning algorithms. The results are listed on the left of Table 4.2 above using the Area Under the receiver operating characteristic Curve (AUC) measure under various positive vs. negative ratios settings. It is clear that MyPHI outperformed all other methods in most cases, and was comparable with LinSVM-W under 1:10 ratio.

In addition, it can be observed that baseline methods, without considering class-weighting, performed poorly at ratio 1:100. On the other hand, the class-weighted versions of baselines (i.e., LinSVM-W and LR-W) was able to better handle class imbalance, though not as good as MyPHI. In fact, MyPHI achieved its highest performance at 89.95% averaged AUC under 1:100 ratio.

On the right of Table 4.2, we compare the algorithms in a noisy case, where the non-target instances were introduced by sampling the same amount as the target instances. It can be seen that the performance was greatly compromised in the case of 1:1 ratio for all algorithms. This shows that non-target cases did confuse learning models given limited learning instances. However, when the portion of negative instances increased, the performance bounced back. This may explain how larger training instances could help mitigate noise. In addition, the gap between MyPHI and other methods was enlarged in the noisy case. In fact, the performance of LinSVM-W and LR-W dropped significantly in the noisy case. These results demonstrate the robustness of our method.

Individual Disease Categories

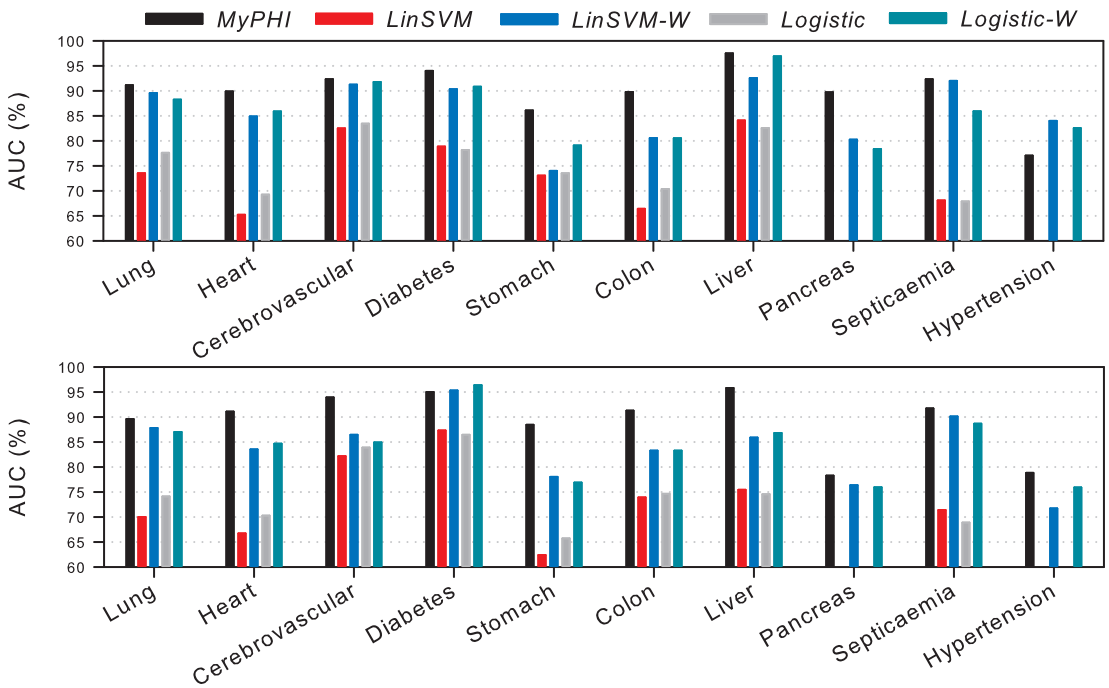

FIGURE 4.2: The AUC for individual disease categories under the 1:100 positive vs. negative ratio. The top figure shows the performance in the clean case, where no instances of non-target disease categories are present. The bottom figure displays the performance in the noisy case, where instances of non-target disease categories are introduced.

We further compared the results at the level of individual disease category. Figure 4.2 shows the performance of the clean case in the top figure and noisy case in the bottom figure. It can be seen

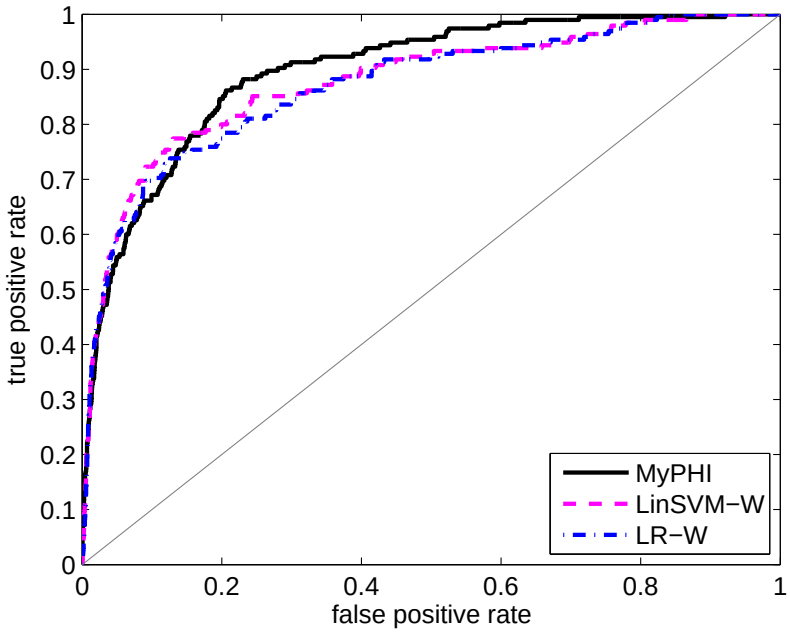

FIGURE 4.3: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 1: Lung).

that, although performance varied according to the categories, MyPHI generally outperformed the other methods. One exception was Hypertension in the clean case; however, our method performed better in the noisy case. Another exception was Diabetes in the noisy case. For Heart, Cerebrovascular, Stomach, Colon, Liver, and Hypertension disease categories, MyPHI showed significantly better results than the other methods.

We looked into the Receiver Operating Characteristic (ROC) curves of the results in individual disease categories. Figures 4.3 to 4.12 compare the performance of MyPHI, LinSVM-W and LR-W in the noisy case under 1:100 positive vs. negative ratio. The disease categories are ordered as before according to their sizes. MyPHI clearly dominated LinSVM-W and LR-W on the ROC graphs in the cases of Heart, Cerebrovascular, Stomach, Colon, Liver, and Septicaemia categories, while MyPHI was comparable to the two for Lung, Diabetes, and Pancreas categories.

Effects of trace-norm

We also investigated the effects of introducing the trace-norm constraint on W and P (Eq. 4.5), where target and alive cases were used to regulate the less certain labels from the non-target cases. As shown in Figure 4.13 in the Discussion Section 4.3.4, below, introducing W and trace-norm improves the

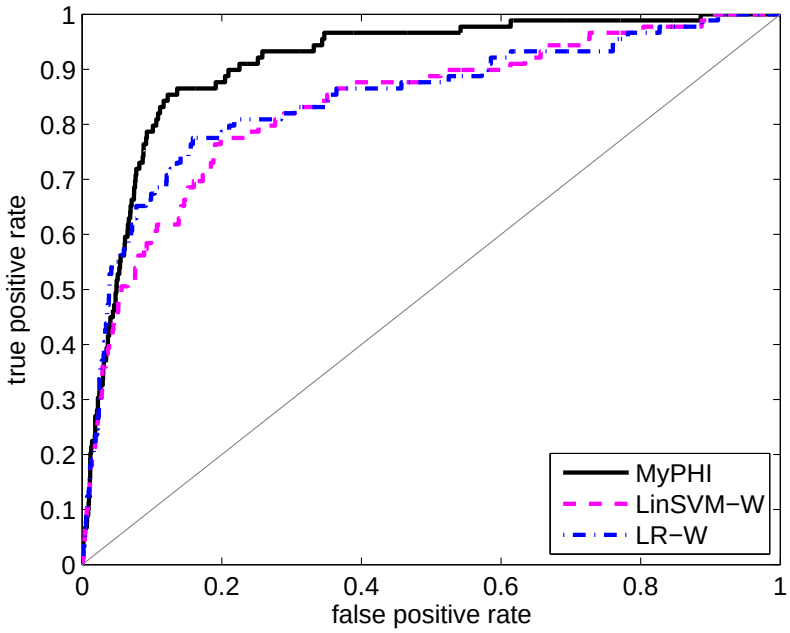

FIGURE 4.4: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 2: Heart).

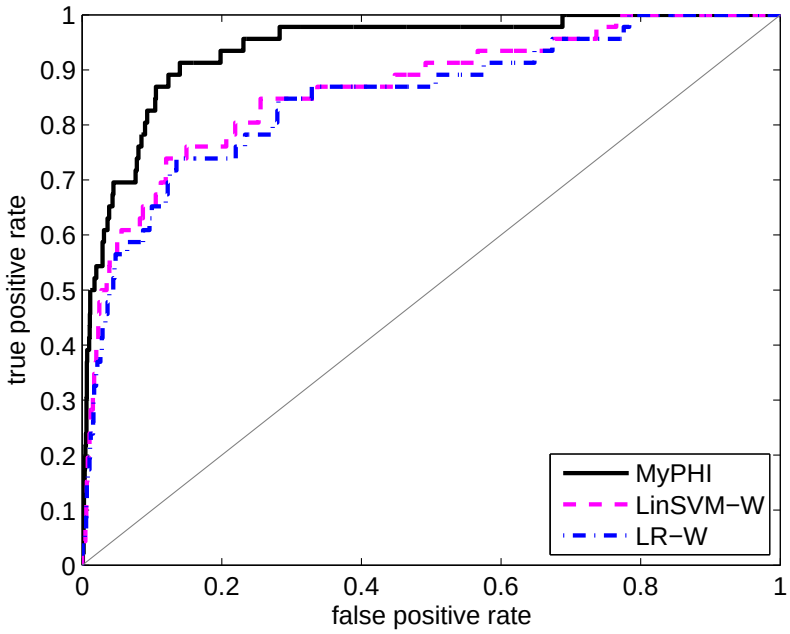

FIGURE 4.5: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 3: Cerebrovascular).

algorithm's performance. The effects were more significant in the cases of positive vs. negative ratio 1:1 and 1:100.

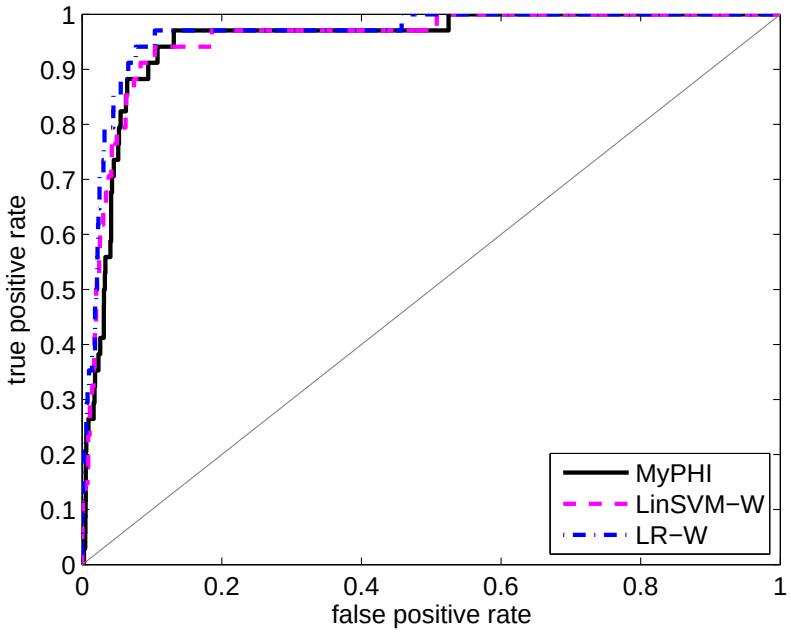

FIGURE 4.6: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 4: Diabetes).

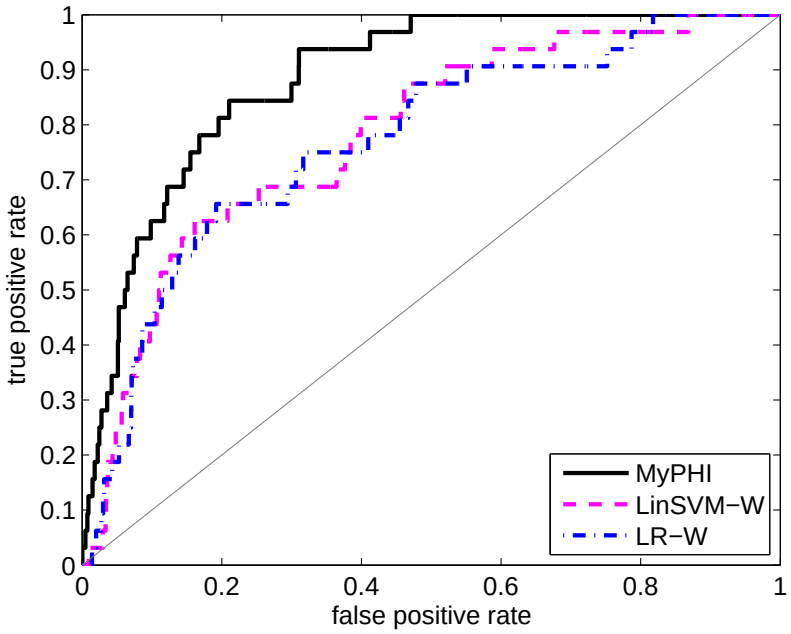

FIGURE 4.7: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 5: Stomach).

Effects of Record Sequence Length

Studies have shown that the incompleteness of data can degenerate prediction performance [57]. As discussed in Section 1.2.1, the averaged record sequence length, i.e., number of records per person,

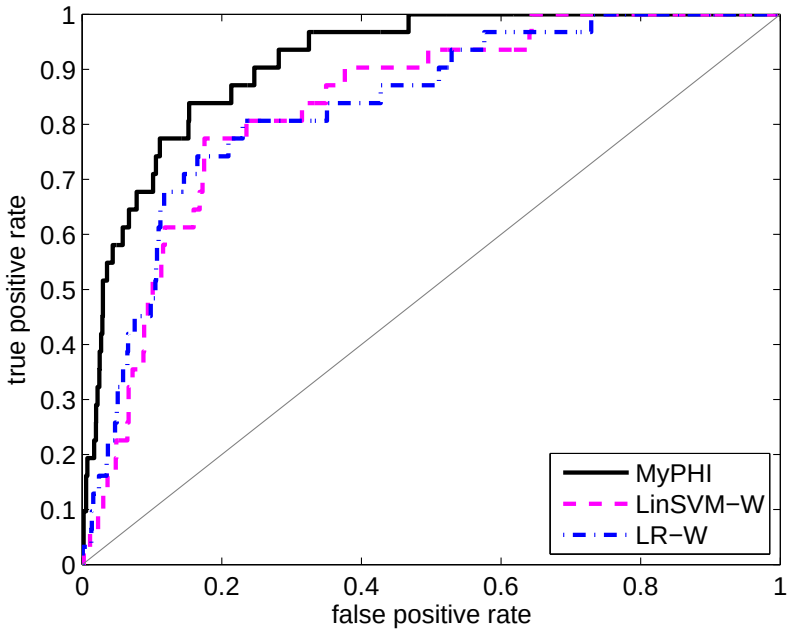

FIGURE 4.8: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 6: Colon).

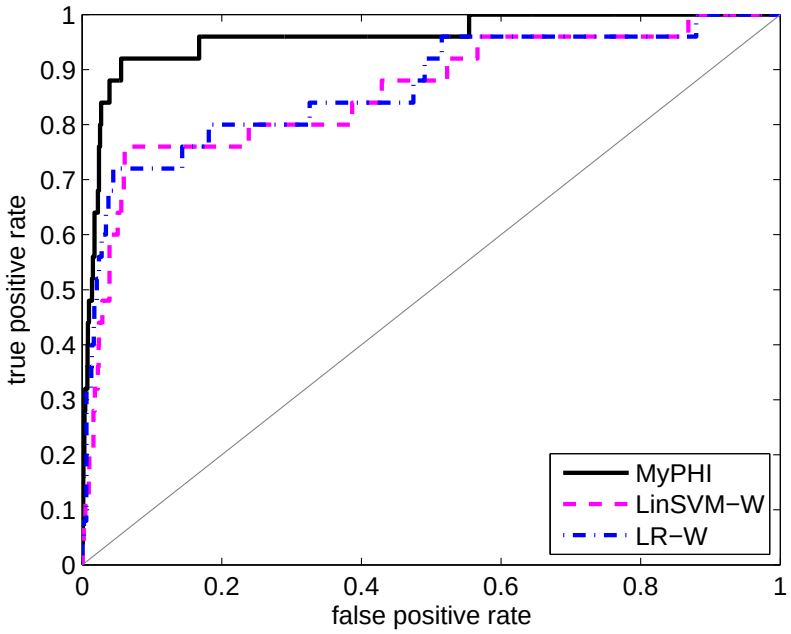

FIGURE 4.9: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 7: Liver).

for our dataset is 2.56, we further conducted experiments to investigate the effects of record sequence length on the performance.

Figure 4.14 shows the averaged AUC under 1:100 ratio with the 95% confidence limits as the error

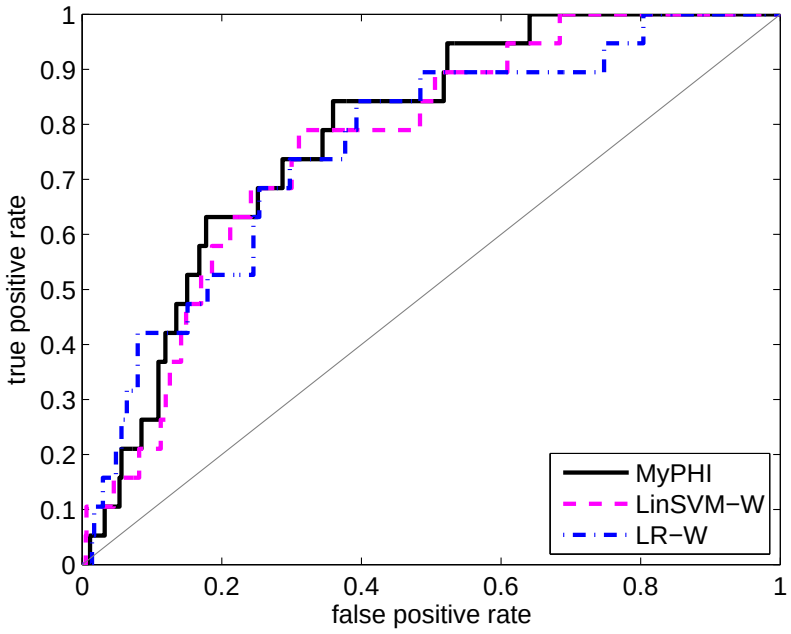

FIGURE 4.10: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 8: Pancreas).

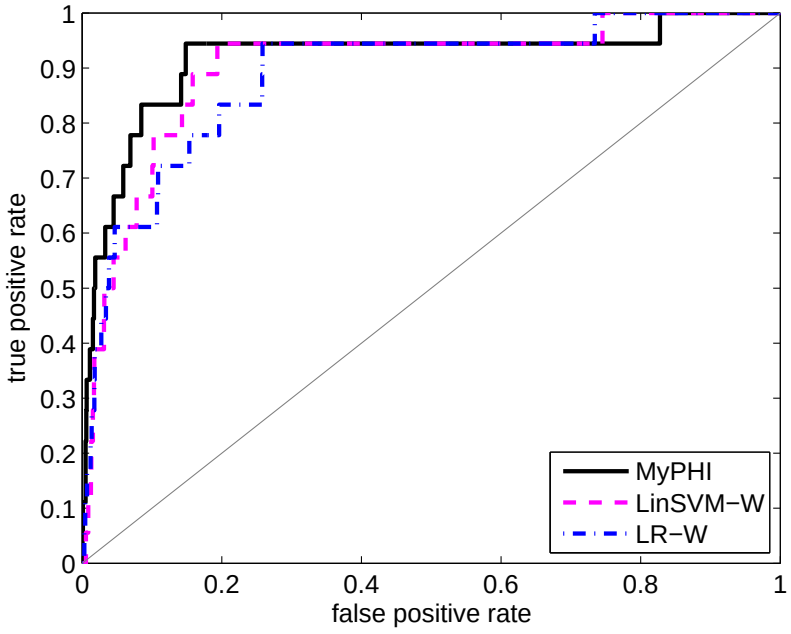

FIGURE 4.11: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 9: Septicaemia).

bars. The Standard Error of the Mean (SEM) is used to calculate the standard error, i.e., $\mathrm{STD}/\sqrt{\mathrm{n_l}}$, where STD is the standard deviation and n_l is the number of cases with record sequence length l. The upper and lower confidence limits can be calculated as $\bar{x} \pm 1.96 \times \mathrm{SEM}$, where $\bar{x}$ is the mean AUC.

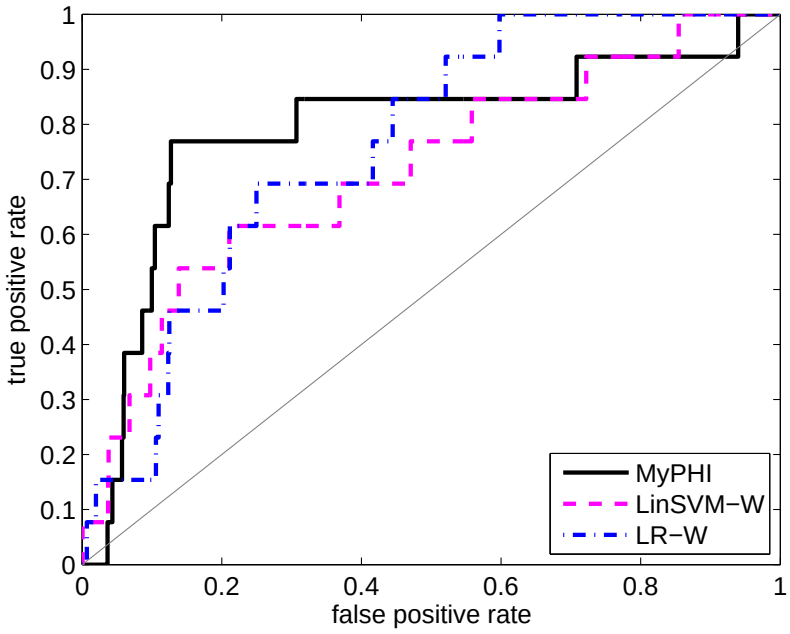

FIGURE 4.12: The ROC of the top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings (Top 10: Hypertension).

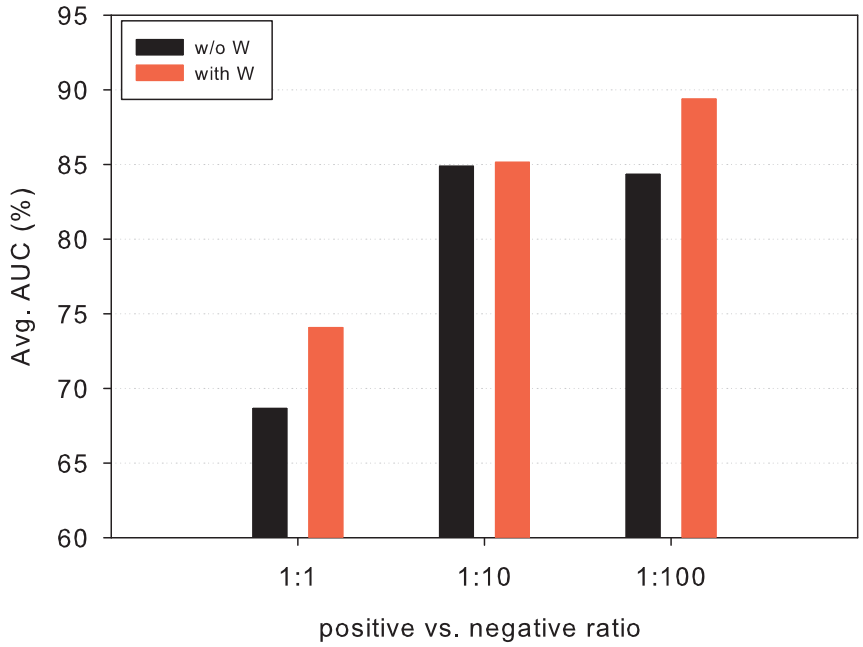

FIGURE 4.13: Comparing the effects of applying the trace-norm in averaged AUC: "with W" denotes Eq. 4.5 where W is introduced and trace-norm applied, while "w/o W" refers to Eq. 4.4 without the effect of trace-norm.

Figure 4.14 shows that predictions with five records had the highest averaged AUC, followed by those with three and four records. Predictions with two records had the lowest averaged AUC score of

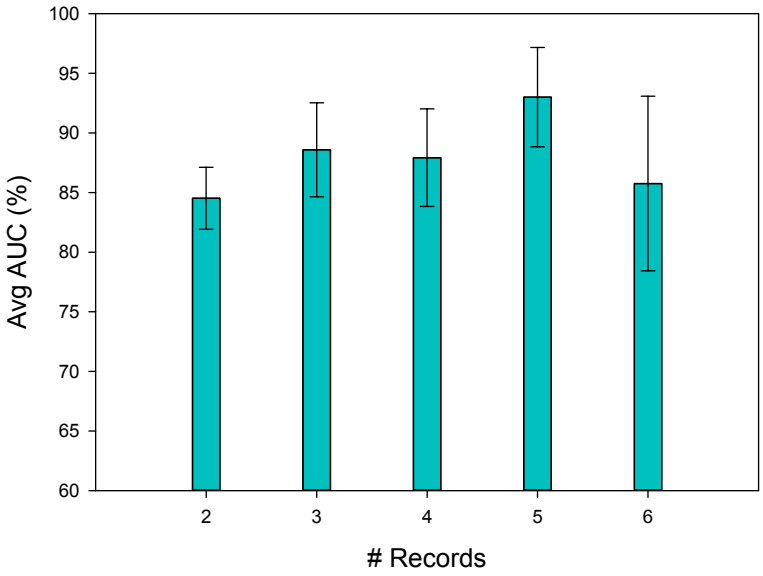

FIGURE 4.14: Comparing the effects of the number of records per person on the performance in averaged AUC. The error bar is calculated by Standard Error of the Mean (SEM).

84.53%, while the averaged AUC for predictions with six records had a larger SEM due to insufficient cases in some runs.

4.3.4 Discussion

Label uncertainty is often observed in real-world medical data. Our extensive experiments on a large GHE dataset with 262,424 check-ups of 102,258 participants have shown the robustness of our model under label uncertainty and class imbalance.

Specifically, MyPHI achieved its best performance at above 89% averaged AUC under 1:100 ratio, about 5% higher than LinSVM-W and LR-W (Table 4.2), two de-facto standard classifiers. The best AUC for a single disease category was achieved in the prediction under the Lung-related disease category at 96.95% AUC (Figure 4.2). The ROC curves (Figure 4.3 and Figure 4.4) further confirmed the above analysis at a fine-grained level. These results suggested that the proposed soft-label learning model is able to better handle label-uncertain data commonly found in medical applications. The parameter S for soft-labeling allowed the algorithm to learn the degree of certainty for every data instance based on the data structure.

In addition, our experiments showed that introducing the terms for W and applying the trace-norm on $E = [W, P]$ helped regularize P, with which the performance is improved (Figure 4.13). This suggested that in dealing with uncertainty, expert knowledge helps.

Finally, we compared the effects of the lengths (2 to 6) of the record sequences on the performance in Figure 4.14. Predictions based on 5 record sequences had the best performance, while those with 2 records had the worst performance. The corresponding confidence limits based on the computed mean and standard error can be taken into account along with the predicted PHI given the record sequence length.

Note that the current model was built based on the GHE dataset, so it is limited to the associated population, namely the elderly residents of Taipei City. However, the proposed methodology is not limited to the current dataset, but can be applied to other medical datasets with similar characteristics as stated in Section 1.2.2. We are also aware of the possible misinterpretation of the predicted results by non-professionals, which could happen to any screening results [51]. So the understanding and interpretation of the predicted PHI scores needs to be assisted by a general practitioner.

4.4 Summary

In this chapter we described MyPHI, a data mining-based method that predicts Personal Health Index (PHI) based on GME records. The extensive experiments on a real-world GME data set of 262,424 records from 102,258 participants demonstrated that our model outperformed the commonly-used classifiers, such as linear SVM, logistic regression and their class-weighted versions. In particular, MyPHI has been shown to be robust under label uncertainty and class imbalance, and achieved 89.95% averaged AUC (Area Under the receiver operating characteristic Curve) under a ratio of 1:100 positive vs. negative.

Chapter 5

Graph-based Health Risk Prediction with Large Unlabeled Data

5.1 Problems and Challenges

General health examination is an integral part of healthcare in many countries. Identifying the participants at risk is important for early warning and preventative intervention. In this study we formulated the task of risk prediction as a multi-class classification problem using the Cause of Death (COD) information as labels, regarding the health-related death as the "highest risk". The goal of risk prediction is to effectively classify 1) whether a health examination participant is at risk, and if yes, 2) predict what the key associated disease category is. In other words, a good risk prediction model should be able to exclude low-risk situations and clearly identify the high-risk situations that are related to some specific diseases.

A fundamental challenge is the large amount of unlabeled data. For example, 92.6% of the 102,258 participants in our geriatric health examination dataset (Section 1.2.1) do not have a COD label. The semantics of such "alive" cases can vary from generally healthy to seriously ill, or anywhere in between. In other words, there is no ground truth available for the "healthy" cases. If we simply treat this set of alive cases as the negative class, it would be a highly noisy majority class. On the other hand, if we take this large alive set as *genuinely unlabeled*, as opposed to cases with known labels removed, it would become a multi-class learning problem with large unlabeled data.

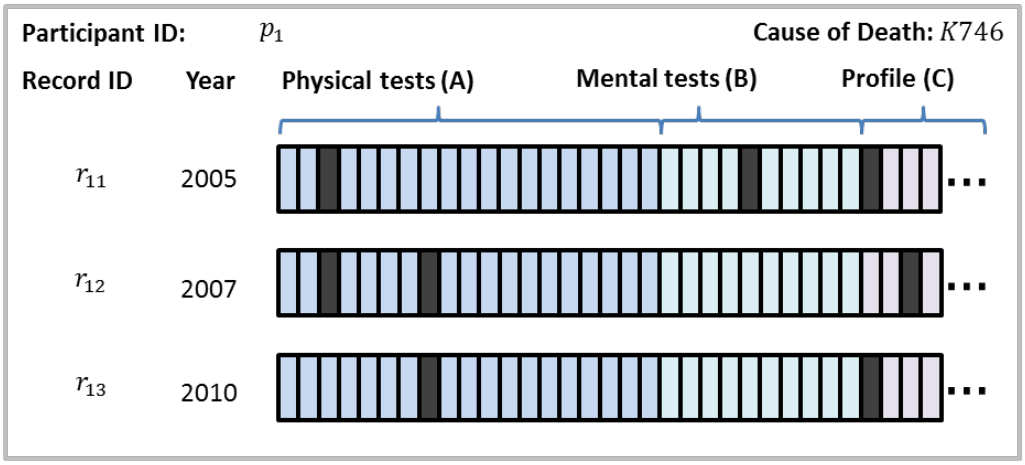

FIGURE 5.1: An example of health examination records of participant p_1 who took examinations in three non-consecutive years, 2005, 2007, and 2010. Test items are in different categories (A,B,...) and the abnormal results are marked black. The main cause of death of p_1 was cirrhosis of liver encoded as $K746$.

The other key challenge of HERs is *heterogeneity*. Figure 5.1 above demonstrates the health examination records of Participant p_1 in three non-consecutive years with test items in different categories (e.g., physical tests, mental tests, etc.) and abnormal results marked black. This example shows that:

- a participant may have a sequence of irregularly time-stamped longitudinal records, each of which is likely to be sparse in terms of abnormal results, and
- test items are naturally in categories, each conveying different semantics and possibly contributing differently in risk identification. Therefore this heterogeneity should be respected in the modeling.

Most existing classification methods on healthcare data do not consider the issue of unlabeled data. They either have expert-defined low-risk or control classes [47, 121, 90, 125, 75, 96] or simply treat non-positive cases as negative [113, 22, 102]. Methods that consider unlabeled data [106, 73, 60, 97, 46, 124, 61, 84, 138] are generally based on Semi-Supervised Learning (SSL) [144] that learns from both labeled and unlabeled data. Amongst these SSL methods, only a few [97, 46] handle large and genuinely unlabeled health data. However, unlike our scenario, both methods are designed for binary classification and have predefined negative cases. A closely related approach is Positive and Unlabeled (PU) learning [83, 130, 131], which can be seen as a special case of SSL with only positive

labels available. While the unlabeled set U in a PU learning problem is similar to our alive set, its existing applications in healthcare only address the binary classification problem. Nguyen *et al.* introduced a multi-class extension called mPUL [95]; however, their method used a combined set of negative and unlabeled examples, while in our case negative examples are not available.

This chapter proposes a semi-supervised heterogeneous graph-based algorithm called **SHG-Health** (Semi-supervised Heterogeneous Graph on Health) as an evidence-based risk prediction approach to mining longitudinal health examination records. To deal with the sparsity of health examination records, it utilizes a graph representation. To handle the *heterogeneity* of health examination records, it explores a heterogeneous graph called **HeteroHER** graph, where examination items in different categories are modeled as different types of nodes and their temporal relationships as links. To tackle large *unlabeled data*, SHG-Health features a semi-supervised learning method that utilizes both labeled and unlabeled instances. In addition, it is able to learn an additional $K+1$ "unknown" class for the participants who do not belong to the K known high-risk disease classes.

The rest of the chapter is organized as follows. Section 5.2 discusses the background on graph-based semi-supervised learning. Section 5.3 presents the proposed SHG-Health algorithm for evidence-based risk prediction. In Section 5.4, we demonstrate the effectiveness and efficiency of our proposed algorithm based on both real datasets and synthetic datasets. Section 5.5 is our summary and findings.

5.2 Background

Learning from labeled and unlabeled data is often called semi-supervised learning or transductive inference [143]. Graph-based methods that model data points as vertices and their relationships as edges on graph, are often used to exploit the intrinsic characteristics of data [52]. Zhu *et al.* [144] proposed an algorithm based on Gaussian fields and harmonic functions to propagate labels to the unlabeled data, which can be interpreted as a random walk on graph. Zhou *et al.* [143] introduced the Learning with Local and Global Consistency (LLGC) algorithm that spreads the label information of each point to its neighbors to achieve both local and global consistency.

A graph can be constructed either 1) based on real-world networked data [116, 68, 66], such as from social networks, bibliographic networks, and web page networks, or 2) by computing affinity

matrices to encode the similarity between data points [143, 99].

Many *graph-based semi-supervised learning* (GSSL) methods can be viewed as estimating a function of soft labels F based on two assumptions on graph [144, 143, 99, 124, 140]. The *smoothness* assumption states that F should not change much for nearby points, and the *fitness* assumption requires that F should not change much from the ground-truth labels. By adapting a graph-based approach and exploring the underlying graph structure of health examination records with semi-supervised learning, our method is capable of handling large unlabeled data.

To further tackle the issues of the absence of ground truth for the "healthy" cases and the heterogeneity embedded in the examination records, we utilized class discovery methods to handle the "unknown" class and heterogeneous graph representations for GSSL as follows.

5.2.1 Class Discovery for Graph-based Semi-supervised Learning

Situations arise when unlabeled data may belong to unknown or latent classes. Nie *et al.* [99] introduced a scholastic graph-based semi-supervised learning (GGSSL) method for novel class discovery (if the number of classes is known) or outlier detection (if otherwise). By introducing an instance-level parameter α that assigns small weight to unlabeled data and large weight to the labeled data, GGSSL allows the soft label scores of unlabeled vertices on the graph to be updated according to their connectedness to labeled vertices.

Wang *et al.* [124] further modified the model to discover more than one unseen class for patient risk stratification based on a patient graph constructed using ICD codes. Recently Zhao *et al.* [140] extended GGSSL for classification on Alzheimer's Disease, by introducing a compact graph construction strategy via minimizing local reconstruction error. However, all of the above algorithms are limited to homogeneous graphs, where vertices belong to one object type, and thus are by themselves not capable of handling the data heterogeneity hidden in health examination records.

To train a disease risk prediction model that is capable of identifying high-risk individuals given no ground truth for "healthy" cases, we treated the "unknown" class as a class to be learned from the known disease classes. We incorporated the class discovery mechanism of [99] into our method to handle the "unknown" class.

5.2.2 Heterogeneous Graph-based Semi-supervised Learning

Traditional GSSL methods have been limited to homogeneous graphs [143, 99, 124, 140]. However, it has been recognized in recent years that networks of heterogeneous types of objects are prevalent in the real world [116, 68, 61, 15]. For example in healthcare applications, methods that explore the heterogeneous structure of gene-phenotype networks have been developed [80, 61]. The term "network medicine" [15] has been coined to refer to a broad approach to human disease based on a complex intracellular and intercellular network that connects tissue and organ systems.

For the heterogeneous extensions of GSSL algorithms, Hwang *et al.* [61] proposed a heterogeneous label propagation algorithm based on GSSL for disease gene discovery. Their heterogeneous disease-gene graph was constructed based on homo-subnetworks that linked same-type objects together and the mutual interactions between homo-subnetworks. The algorithm iteratively propagates the label scores via homo-subnetworks and hetero-subnetworks until convergence.

Ji *et al.* proposed GNetMine [68] to work on a heterogeneous graph of multi-type objects, known as a heterogeneous information network [116]. This classification process can be intuitively viewed as a process of knowledge propagation throughout the network across different types of objects through links. GNetMine was originally designed for bibliographic information networks that were intrinsically heterogeneous and was shown to outperform other GSSL methods with homogeneous graphs.

They further proposed RankClass [66] based on the same framework with additional updates on the local weighted graph for individual classes. However, the above methods were designed for a multi-class semi-supervised learning problem with predefined classes, and thus had no mechanism for handling the "unknown" class. Inspired by GNetMine and RankClass, we integrated a heterogeneous component into our method to handle heterogeneity.

In summary, our proposed SHG-Health algorithm can be seen as combining the advantages of GGSSL [99] and GNetMine [68] for solving a practical clinical problem of risk prediction from longitudinal health examination data with heterogeneity and large unlabeled data issues.

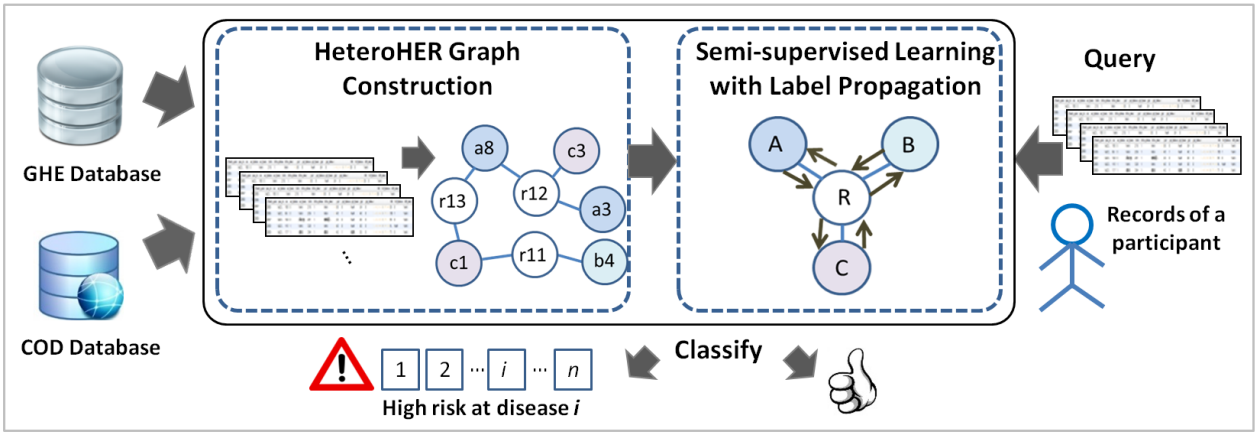

FIGURE 5.2: An overview of the proposed SHG-Health algorithm for risk prediction

5.3 SHG-Health: a Semi-supervised Heterogeneous Graph-based Algorithm for Health Risk Prediction

To solve the problem of health risk prediction based on health examination records with heterogeneity and large unlabeled data issues, we present a semi-supervised heterogeneous graph-based algorithm called SHG-Health. The semi-supervised learning problem is formulated as follows:

Problem 5.1. *Given a set of health examination records of n participants $S = \{s_1, ..., s_l, s_{l+1}, ..., s_n\}$, where $s_i = \{r_{i1}, ..., r_{in_i}\}$ is the set of n_i records of participant p_i and r_{ij} is a tuple (x_{ij}, t_{ij}) such that $x_{ij} \in \mathbb{R}^d$ is a d-dimensional vector for the observations at time t_{ij}, and a set of labels $C = \{1, ..., c\}$, the first l participants s_i $(i \leq l)$ are labeled as $y_i \in C$ and the remaining $u = n - l$ participants $s_{l+1}, ..., s_{l+u}$ are unlabeled ($l \ll u$). The goal is to predict for unlabeled $s_i (l < i \leq n)$ a label $y_i \in \tilde{C} = \{1, ..., c, c+1\}$ where $c+1$ gives a mechanism to handle an additional class for unknown cases.*

An overview of our proposed solution to the problem is included in Figure 5.2, above. Our SHG-Health algorithm takes General Health Examination data (GHE) and the linked Cause of Death (COD) labels described in Section 1.2.1 as inputs. Its key components are a process of *Hetero*geneous Health Examination Record (HeteroHER) graph construction and a semi-supervised learning mechanism with label propagation for model training. Given the records of a participant p_i as a query, SHG-Health predicts whether p_i falls into any of the high-risk disease categories or "unknown" class whose instances do not share the key traits of the known instances belonging to a high-risk disease class.

5.3.1 HeteroHER Graph

A graph representation allows us to model data that is sparse. It is useful for modeling data such as our GHE records, where abnormal results are of interest. To capture the heterogeneity naturally found in health examination items, we constructed a graph called HeteroHER consisting of multi-type nodes based on health examination records.

Graph Construction

The process of HeteroHER graph construction included the following steps:

Step 1. ***Binarization:***

As a preparatory step, all the record values were first discretized and converted into a $0/1$ *binary representation, which served as a vector of indicators for the absence/presence of a discretized value. Specifically, real values, such as age, were first binned into fixed intervals (e.g., 5 years). Then, all the ordinal and categorical values were converted into binary representations.*

Step 2. ***Node Insertion:***

Every element in the binary representation obtained in Step 1 with a value "1*" was modeled as a node in our HeteroHER graph, except that only the abnormal results were modeled for examination items (both physical and mental). This setting was primarily based on the observation that physicians make clinical judgements generally based on the reported symptoms and observed signs, and secondarily for the reduction of graph density.*

Step 3. ***Node Typing:***

Every node was typed according to the examination category that its original value belonged to, for example, the Physical tests (A), Mental tests (B), and Profile (C) in Figure 5.1. In addition, a new type of nodes was introduced to represent individual records such as r_{11}*,* r_{12}*, and* r_{13} *in the same figure.*

All the other non-Record type nodes that were linked to the Record type nodes could be seen as the attribute nodes of these Record type nodes. In other words, categories A, B, and C in Figure 5.1 could be regarded as the attributes of the Record type at a schema level. This leads to a graph schema

with a star shape as shown on the right of Figure 5.3, which is known as a star schema [116]. Note that types can often be hierarchically structured and thus choosing the granularity of node type may require domain knowledge or be done experimentally.

Step 4. ***Link Insertion:***

Every attribute (non-Record) type node was then linked to a Record type node representing the record that the observation was originally from. The weight of the links was calculated based on the assumption that the newer a record the more important it was in terms of risk prediction. A simple function $g(\cdot)$ can be defined as:

$$g(t) = \frac{t - s + 1}{1} \tag{5.1}$$

where t is the time of current record, l is the time window of interest, and s is the starting time of the time window.

Other functions such as truncated Gaussian distribution and Chi Squared distribution could also be used [23]. The window length can be selected according to the application needs. For example, the time unit for our risk prediction is in years and the window length was set as 6 years.

We included Figure 5.3 as an example based on the records of participant p_1 in Figure 5.1 to illustrate the process. In this simplified example, we assumed all the values of examination items were binary. Different types of examination items in Figure 5.1 were treated as different types of nodes on the graph. An abnormal result of the i^{th} item of type Z in the j^{th} record of the k^{th} participant is represented as a link between nodes r_{kj} and z_i. For instance, there is a link between r_{11} and a_3 in the left sub-figure of Figure 5.3, and the weight of the link is $(2005 - 2005 + 1)/6 \cong 0.17$ using Eq. (5.1) with a window width equal to 6 years.

The output of the graph construction process was a heterogeneous graph represented as a set W of sparse matrices W_{ij} for any two node types i, j that are linked to each other in the schema on the right in Figure 5.3.

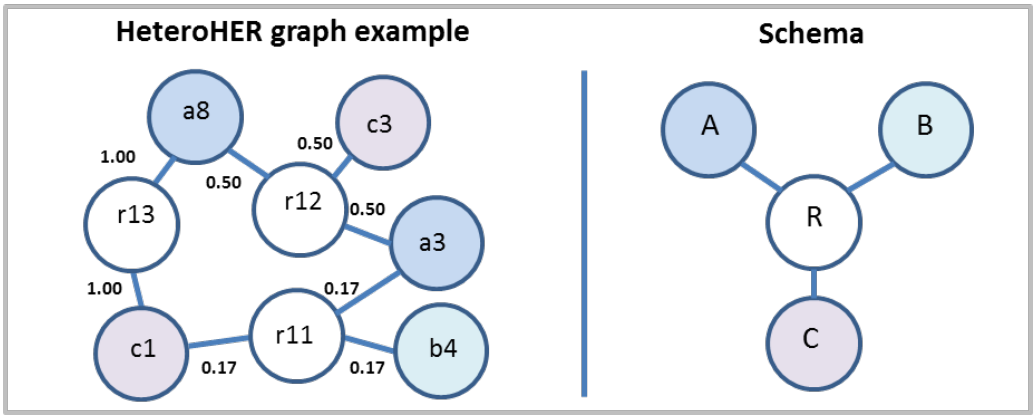

FIGURE 5.3: The graph on the left shows a HeteroHER graph extracted from the example in Figure 5.1. For instance, there is a link between r_{11} (the first record of p_1) and a_3 (the third item of category A) if the result of a_3 is abnormal in r_{11}. The link is weighted using Eq. (5.1). The star-shaped schema on the right is a type-level schema of such a graph.

Normalized Weights

To strengthen the weights in the low density region and weaken the weights in the high density region, the weights W_{ij} for $i, j = 1, ..., m$ were further normalized by the row sum and column sum as in [99]:

$$\tilde{W}_{ij} = D_{ij}^{-1/2} W_{ij} D_{ji}^{-1/2} \tag{5.2}$$

where D_{ij} is an n_i-by-n_i diagonal matrix with the (p, p) element as $d_{ij,pp}$ and $d_{ij,pp} = \sum_q W_{ij,pq}$ is the sum of row p in W_{ij}.

5.3.2 Semi-supervised Learning on HeteroHER Graph

The second component of our method was a semi-supervised learning algorithm for the constructed HeteroHER graph (Section 5.3.1). The algorithm combines the advantages of [99] for class discovery and [68] for handling heterogeneity to solve a specific problem induced by evidence-based risk prediction from health examination records.

In this section, we first define an objective function for the learning problem and show its convexity, followed by an optimization procedure to solve the problem. Then we derive an efficient iterative algorithm and show its convergence. Finally, time complexity is discussed.

Notations

Let us start with definitions and notations for the following discussions. Assume there are c classes and there is one additional "unknown" class for the cases that are not known to belong to any of the c disease classes. In this work we attach label information of a participant to the Record type nodes representing their examination records. However, the model is general enough to include labels for different types of nodes.

Defining $Y = [Y_1, ..., Y_m]^{\mathrm{T}} \in \{0,1\}^{\sum_i^m n_i \times (c+1)}$ such that $Y_i = [y_{i1}, ..., y_{in_i}]^{\mathrm{T}} \in \{0,1\}^{n_i \times (c+1)}$ encodes the labels of type i nodes. Let $y_{ip}^{(k)}$ be the k^{th} element of vector y_{ip}. If x_{ip}, i.e., node p of type i, is labeled, $y_{ip}^{(k)} = 1$ if x_{ip} belongs to class k; otherwise $y_{ip}^{(k)} = 0$. If x_{ip} is unlabeled, $y_{ip}^{(c+1)} = 1$. By doing so, we set the initial labels of the unlabeled data to be the unknown class. However, we will show later (Section 5.3.3) that these initial labels for the unlabeled data have little influence on learning their labels.

In addition, we designed the computed labels to be soft labels. Soft labels are especially desirable for medical applications because they provide a level of confidence for the prediction. Let $F = [F_1, ..., F_m]^{\mathrm{T}} \in \mathbb{R}^{\sum_i^m n_i \times (c+1)}$ be the computed soft labels of m node types such that $F_{ip} \in \mathbb{R}^{c+1}$ is a vector indicating the degree of certainty that x_{ip} belongs to any of the $c+1$ classes. The class label of x_{ip} is computed as:

$$\arg \max_{k \leq (c+1)} F_{ip}^{(k)}$$

and F_i can be initialized uniformly amongst type i nodes for $i = 1, ..., m$.

Objective Function

We considered a regularized framework on a heterogeneous graph for our problem, denoting $tr(\cdot)$ as trace and $\|\cdot\|_F$ as the Frobenius norm of matrix, i.e., $\|M\|_F^2 = tr(M^T M)$. The classification problem can then be viewed as an optimization problem that minimizes an objective function $J(F)$:

$$J(F) = \sum_{ij}^{m} \gamma_{ij} \sum_{p}^{n_i} \sum_{q}^{n_j} \tilde{W}_{ij,pq} \left\| F_{ip} - F_{jq} \right\|_F^2 + \sum_{i}^{m} \sum_{p}^{n_i} \mu_{ip} \tilde{d}_{ip} \left\| F_{ip} - Y_{ip} \right\|_F^2 \tag{5.3}$$

where $\tilde{W}_{ij}$ is the normalized weights on the links between type i and j nodes as defined in Eq. (5.2), above, and F and Y are the same as defined earlier.

The first term is the *smoothness* constraint based on the assumption that if two nodes in a graph are close to each other, their computed label values should be close. Let $z = [z_1, ..., z_m]^T$ where $0 \leq z_i \leq 1$ be the weights for m node types. Then, γ_{ij} is defined as between-type weight of type i and type j nodes as follows:

$$\gamma_{ij} = \begin{cases} \frac{1}{2} z_j & \text{if } i = j \\ z_j & \text{otherwise} \end{cases} \tag{5.4}$$

where the weight is reduced to half for links between same-type nodes to avoid double counting in the summation.

The second term is the *fitness* constraint that penalizes when the computed F is different from labels Y. Let $\tilde{d}_{ij,pp} = \sum_q \tilde{W}_{ij,pq}$ and $\tilde{D}_{ij}$ be an n_i-by-n_i diagonal matrix with the (p, p) element as $\tilde{d}_{ij,pp}$. Then $\tilde{d}_{ip}$ in Eq. (5.3) is the degree of node p of type i, weighted by its connected node type, i.e., $\tilde{d}_{ip} = \sum_j^m \sum_p^{n_i} z_j \tilde{D}_{ij,pp}$. On the other hand, $\mu_{ip} > 0$ controls the trade-off between the smoothness and fitness constraints for node p of type i.

Convexity

The strict convexity of Eq. (5.3) minimization is derived in this section. For clarity, we will discuss two terms in Eq. (5.3) separately. The first term of the objective function denoted as $J_1(F)$ can be

derived as follows:

$$\begin{aligned} J_1(F) =& \sum_{i}^{m}(\sum_{j}^{m}\gamma_{ij}\sum_{p}^{n_i}\sum_{q}^{n_j}\tilde{W}_{ij,pq}\left\|F_{ip}-F_{jq}\right\|_F^2) \\ =& \sum_{i}^{m}\sum_{j}^{m}\gamma_{ij}tr(F_i^T\tilde{D}_{ij}F_i - 2F_i^T\tilde{W}_{ij}F_j + F_j^T\tilde{D}_{ji}F_i) \\ =& tr(\sum_{i}^{m}\sum_{i\neq j}^{m}z_j(F_i^T\tilde{D}_{ij}F_i - 2F_i^T\tilde{W}_{ij}F_j + F_j^T\tilde{D}_{ji}F_i) \\ & + z_iF_i^T(\tilde{D}_{ii}-\tilde{W}_{ii})F_i) \end{aligned} \tag{5.5}$$

where $\tilde{W}_{ij}$ is defined in the same way as in Eq. (5.2) and $\tilde{D}_{ij}$ and γ_{ij} the same as in Eq. (5.3).

Suppose the total number of nodes $n = \sum_i^m n_i$. Let $\tilde{L}$ be a n-by-n block matrix. Let its (i,j) block $\tilde{L}_{ij} = \tilde{D}_{ij} - \tilde{W}_{ij}$ be a Laplacian matrix with normalized weights. Eq. (5.5) can be transformed to the following matrix expression:

$$J_1(F) = tr(F^T I_z \tilde{L} F) = tr(F^T H F) \tag{5.6}$$

where I_z is a block diagonal matrix with the elements of the diagonal of (i,i) block equal to z_i, and $H = I_z\tilde{L}$.

The second term of the objective function can be derived as follows:

$$\begin{aligned} J_2(F) &= \sum_{i}^{m}\sum_{p}^{n_i}\mu_{ip}\tilde{d}_{ip}\left\|F_{ip}-Y_{ip}\right\|_F^2 \\ &= tr(\sum_{i}^{m}(F_i-Y_i)^T U_i\tilde{D}_i(F_i-Y_i)) \end{aligned} \tag{5.7}$$

where the diagonal matrix $\tilde{D}_i = \sum_j^m z_j\tilde{D}_{ij}$ and its (p,p) entry is $\tilde{d}_{ip}$ as defined earlier in Section 5.3.2. U_i is a diagonal matrix such that $U_{i,pp} = \mu_{ip}$.

Combining Eq. (5.6) and Eq. (5.7), the objective function can be transformed into:

$$J(F) = tr(F^T H F) + tr((F-Y)^T U\tilde{D}(F-Y)) \tag{5.8}$$

where $\tilde{D}$ is a block diagonal matrix with the diagonal of the (i, i) block equal to $\tilde{D}_i$.

It is easy to verify that $\tilde{L}$ is positive semi-definite and likewise, H, U, $\tilde{D}$, $U\tilde{D}$ and their traces. Therefore the objective function is strictly convex.

Optimization Procedure

The closed-form solution for minimizing Eq. (5.8) can be obtained by setting the partial derivative of $J(F)$ with respect to F to zero:

$$\left.\frac{\partial J(F)}{\partial F}\right|_{F=F^*} = 2(HF^* + \tilde{D}U(F^* - Y)) = 0 \tag{5.9}$$

where we use the fact that $H = I_z\tilde{L}$ is symmetrical and that $\tilde{D}$ and U are diagonal.

By multiplying $\tilde{D}^{-1}$ on both sides of Eq. (5.9) and rearranging the equation, we have:

$$F^* = (\tilde{D}^{-1}I_z\tilde{L} + U)^{-1}UY \tag{5.10}$$

By using the fact that $U = \frac{I-I_\alpha}{I_\alpha} = \frac{I_\beta}{I_\alpha}$, we have:

$$\begin{aligned} F^* &= (\tilde{D}^{-1}I_z\tilde{L}I_\alpha + I_\beta)^{-1}I_\beta Y \\ &= (I_\alpha(I - I_zP) + I_\beta)^{-1}I_\beta Y \\ &= (I - I_\alpha I_z P)^{-1}I_\beta Y \\ &= (I - \hat{P})^{-1}I_\beta Y \end{aligned} \tag{5.11}$$

where $\tilde{L} = \tilde{D} - \tilde{W}$ as before, $P = \tilde{D}^{-1}\tilde{W}$, and $\hat{P} = I_\alpha I_z P$.

Note that the $\infty-$norm of $I_\alpha I_z P$ is lower than 1 given $0 \leq z_i\alpha_i < 1(i = 1, ..., n)$. Hence the spectral radius of $\hat{P}$ is not greater than the $\infty-$norm. So $(I - \hat{P})$ is invertible.

5.3.3 Iterative Solution

An iterative algorithm is often more efficient than a closed-form solution with the inverse of a matrix. Here we describe an iterative solution for F and prove its convergence. The optimal F_i for type i

nodes can be computed with the following update rule for $i = 1, ..., m$:

$$F_i(t+1) = I_{\alpha_i} \sum_{j}^{m} z_j P_{ij} F_j(t) + I_{\beta_i} Y_i \tag{5.12}$$

where $P_{ij} = \tilde{D}_i \tilde{W}_{ij}$, $I_{\beta_i} = I - I_{\alpha_i}$, and z_j is a type weight scalar as defined earlier in Section 5.3.2.

Eq. (5.12) bears a label propagation interpretation. Each node iteratively spreads label information to its neighbors until a global stable state is achieved. Particularly, $z_j P_{ij}$ can be seen as the normalized links from type j nodes to type i nodes, scaled by the source type weight z_j. Soft labels of type i nodes at $(t+1)$ are determined by two factors, 1) the computed label scores of neighboring nodes at time t propagated via links, and 2) the initial labels for type i nodes. The diagonal matrix I_{α_i} controls the trade-off between these two influences.

It is important to note that I_{α_i} provides a mechanism to learn an extra outlier class via an instance-level control over the trade-off. Specifically, α_l and α_u are introduced as parameters in the range of $[0, 1]$ that control the influence from the labeled data and unlabeled data respectively.

The parameters α_{ip} $(p = 1, ..., n_i)$ in I_{α_i} are defined as $\alpha_{ip} = \alpha_l$ if x_{ip} is labeled, and $\alpha_{ip} = \alpha_u$ if otherwise. The larger α_l and α_u are, the less influence initial labels from Y has. Particularly, when α_u is set to a value extremely close to 1, it means that the initial labels of the unlabeled data play almost no role in the learning so that the computed label for an unlabeled case is basically determined by its connectivity in the graph. This mechanism allows the algorithm to learn an additional $(c+1)$ class for nodes that are less connected to the labeled nodes from high risk disease classes.

The complete algorithm of SHG-Health, combining the graph construction and iterative solution, is summarized in Algorithm 3.

Convergence

The proof of the convergence of Eq. (5.12) is as follows: Let I_α, I_β, z_j and $P = \hat{D}^{-1}\hat{W}$ be the same as defined earlier in Section 5.3.2. The update rule Eq. (5.12) for type $i = 1, ..., m$ can be reorganized

Algorithm 3 SHG-Health

Input: a set of health examination records of n participants S, the corresponding encoded labels Y
Output: optimized F as the computed soft labels

1: $W \leftarrow$ graph construction from S (Section 5.3.1).
2: Calculate the normalized weights for $i, j = 1, ..., m$ by: $\tilde{W}_{ij} = D_{ij}^{-1/2} W_{ij} D_{ji}^{-1/2}$ (5.2)
3: Initialize F_i uniformly amongst type i nodes for $i = 1, ..., m$.
4: $t = 1$
5: **repeat**
6: Update F_i for $i = 1, ..., m$ by:
7: $F_i(t+1) = I_{\alpha_i} \sum_{j}^{m} z_j P_{ij} F_j(t) + I_{\beta_i} Y_i$ (5.12)
8: $t = t + 1$
9: **until** convergence **return** F

as:

$$\begin{aligned} F_i(t+1) &= I_{\alpha_i} \sum_{j}^{m} z_j P_{ij} F_j(t) + I_{\beta_i} Y_i \\ &= I_{\alpha_i} P_i I_{z_i} F(t) + I_{\beta_i} Y_i \\ &= \hat{P}_i F(t) + I_{\beta_i} Y_i \end{aligned} \tag{5.13}$$

where $\hat{P}_i = I_{z_i} I_{\alpha_i} P_i$. It is equivalent to the following expression:

$$F(t+1) = \hat{P} F(t) + I_\beta Y \tag{5.14}$$

It has been proven in [99] that $F^* = \lim_{t \to \infty} F(t) = (I - \hat{P})^{-1} I_\beta Y$, which is equivalent to the close-form solution expressed in Eq. (5.11), and hence our proof is completed.

Time Complexity Analysis

Now we analyze the computational time complexity of the iterative solution (Step 3-8 of Algorithm 3). Step 3 takes $O(k|V|)$ time for initialization, where k is the number of classes (i.e., $k = c + 1$) and $|V|$ is the number of nodes in the graph. At each iteration of Step 6, every link needs to be processed twice, once for the node at each end of the link. This is done for every class, and consequently takes $O(k|E|)$ time, where $|E|$ is the number of links. Also, another $O(k|V|)$ time is needed for incorporating $I_{\beta_i} Y_i$. Therefore, the total time for each iteration is $O(k(|E| + |V|))$, and the total time

complexity for the entire iterative solution is $O(lk(|E| + |V|))$, where l is the number of iterations.

5.4 Experiments and Results Analysis

Synthetic Datasets

To test the stability of our method, we generated two groups of synthetic datasets based on the distribution of the processed real datasets. Specifically, we computed the value distributions for a given disease class, a given year, and a given feature for 10 disease classes selected based on COD codes. The selection of these disease classes is reported in Section 5.4.1.

The cumulative distribution functions (CDFs) were computed, based on which a random number generator was used to select a value for a given disease class, a year, and a feature. The resulting synthetic datasets are as follow.

- **Balanced datasets**. To test the stability with increasing class size in a balanced-class setting, we generated the same number of instances for all disease classes and the "unknown" (unlabeled) class with increasing class size in the range of $\{100, 300, 500, 1000\}$.

- **Increasing unlabeled size datasets**. To test the stability, given the increasing scale of unlabeled data, we generated 1000 instances for every disease class, with the "unknown" (unlabeled) class size equivalent to $1000 \times p$, where p is a scalar in the range of $\{1, 3, 5, 10, 15, 20\}$.

Note that the synthetic datasets were generated based on the assumption that features were independent, which does not hold in real-world healthcare data. However, by using synthetic datasets we could better understand algorithm behaviors in different class settings.

5.4.1 Experimental Settings

High-risk Disease Classes

We selected 10 ICD disease categories as high-risk disease classes. The first 3 digits of ICD 10 were used to define disease categories and mapped to the corresponding 3-digit ICD 9 codes for the

TABLE 5.1: Sizes of 10 disease categories and unlabeled cases in terms of participant (P) and record (R).

Class	ICD10	Name	ICD9	Size (P)	Size (R)
1	J18	Pneumonia, organism unspecified	481,485,486,514	434	851
2	J40-J44	Chronic lower respiratory diseases	490-493,496	277	531
3	C34	Malignant neoplasm of bronchus and lung	162	253	586
4	C22	Malignant neoplasm of liver and intrahepatic bile ducts	155	219	375
5	C16	Malignant neoplasm of Stomach	151	193	355
6	C18	Malignant neoplasm of colon	153	186	371
7	I25	Chronic ischaemic heart disease	412,414,429	185	330
8	C25	Malignant neoplasm of pancreas	157	123	242
9	C23	Malignant neoplasm of gallbladder	156	74	126
10	G20	Parkinson's disease	332	37	73
Unlabeled (alive) cases				26,771	69,802

records taken before 2009. Table 5.1, shows the size of the 10 disease classes in terms of the number of participants and number of records.

The process of selecting 10 disease categories is as follows. We selected diseases with top-10 frequency counts. Based on suggestions from clinical experts, we excluded diabetes mellitus, which was known to have many complications, and acute myocardial infarction, an acute disease. This gave us the first 8 classes in Table 5.1.

To make the problem closer to a real-world situation, we selected two additional diseases with less frequency to represent minority classes. The unlabeled cases were then randomly selected from the alive cases with a ratio equal to data distribution i.e., the deceased/alive ratio as 7.4/92.6. This gave us 69,802 records from 26,771 unlabeled participants (Table 5.1).

HeteroHER Graph Construction

Four node types were modeled for constructing our HeteroHER graph, namely *Record*, *Physical Test*, *Mental Test*, and *Profile*. *Physical Test* refers to all the lab tests and physical examinations in Table 1.1 introduced earlier in Chapter 1, while *Mental Test* covers both the mental health assessments and cognitive function assessments. *Profile* includes all the patient demographics and habits, and *Record* indicates the artificial nodes created for representing individual records. As discussed earlier, only the abnormal results, both from physical and mental tests, were included in graph construction.

Table 5.2 shows the network statistics of the HeteroHER graphs extracted from both the real and

TABLE 5.2: Extracted HeteroHER graph statistics

Dataset		# participant	# nodes		# links to Record nodes				Density
			Record	Total	Test	Mental	Profile	Total	
Real	GHE@10class	26,771	73,642	73,778	601,062	119,952	523,387	1,244,401	0.1242
Synthetic	(100,100)	1,100	3,013	3,149	28,071	4,611	22,552	55,234	0.1348
	(300,300)	3,300	9,054	9,190	84,052	13,982	68,053	166,087	0.1349
	(500,500)	5,500	15,092	15,228	139,736	23,134	113,231	276,101	0.1345
	(1000,1000)	11,000	30,201	30,337	280,412	46,655	227,932	554,999	0.1351
	(1000,3000)	13,000	35,463	35,599	323,101	55,263	265,067	643,431	0.1334
	(1000,5000)	15,000	40,695	40,831	365,005	63,974	301,661	730,640	0.1320
	(1000,10000)	20,000	53,674	53,810	469,575	85,167	393,486	948,228	0.1299
	(1000,15000)	25,000	66,841	66,977	575,360	106,694	485,914	1,167,968	0.1285
	(1000,20000)	30,000	79,979	80,115	682,022	128,357	579,269	1,389,648	0.1278

synthetic datasets. Note that density was calculated as the ratio of the number of edges E to the number of possible edges P. Since HeteroHER Graph has a star schema (Figure 5.3), P is calculated as the number of *Records* nodes times the number of all attribute type nodes.

Evaluation Metrics

We designed a two-stage evaluation strategy to evaluate the proposed SHG-Health.

In the first stage, we evaluated an algorithm's ability to identify high-risk cases, regardless of their disease category. All predicted disease cases were regarded as *predicted positive* cases and the true disease cases were regarded as (true) *positive* cases. As there was no ground truth for the negative or healthy cases, we used measures that focused on positive predictions, namely precision, recall/sensitivity, and F-score. While precision measures how accurate the positive predictions of an algorithm are, recall shows its ability to catch the positive cases. F-score calculates the harmonic mean between precision and recall.

In the second stage, we looked into a method's ability to predict the correct disease class given that it predicted a case to be in one of the high-risk classes. This was a conditional evaluation that only considered cases that were predicted as one of the disease classes. Macro-precision and macro-recall measures were used. Macro-averaging took the average of precision or recall scores computed from individual classes [79]. It assumed that all classes were equally important, so that the performance of minority classes could be reflected in the macro-averaged scores.

Algorithms for Comparison

For SHG-Health, we compared three time-weighted functions for graph construction, namely Eq. (5.1) (SHG-Health), truncated Gaussian (SHG-Health-Gaus), and truncated Chi Squared (SHG-Health-Chi2) discussed in Section 5.3.1. Other algorithms that were compared with our method were:

- **Support Vector Machines:** SVM was adopted as one of our baseline methods. Although SVM with RBF kernel achieved the best results in [22], linear and RBF kernels had very similar performance in our experiments. We only report the results of the linear kernel for its favourable efficiency. The LIBSVM [18] and LIBLINEAR [39] implementations were used in our experiments for the RBF and linear kernels respectively.

- **Nearest Neighbor Classifier**: KNN classifier is a common baseline for graph-based models. K was experimentally set to 1.

- **General Graph-based Semi-Supervised Learning:** GGSSL [99] is a state-of-the-art graph-based semi-supervised method for class discovery. As it is not directly applicable to heterogeneous graphs, we constructed a homogeneous graph by converting all types of nodes in our heterogeneous graph into a single-type graph. The MATLAB implementation available from the author web page [98] was employed in our experiments.

- **GNetMine:** GNetMine [68] is a state-of-the-art graph-based semi-supervised method on a graph of heterogeneous nodes. The MATLAB implementations by the authors, available from GitHub [65], was employed in our experiments.

For parameter tuning, the parameters c and γ for SVM were tuned based on the $\{10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$ grid. The parameters α_i and λ_{ij} in GNetMine denote type weights and type relationship weights respectively. They were tuned based on the α_i/λ_{ij} ratio grid $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ with α_i fixed to 0.1 as in [68]. For fair comparison to GNetMine, the type weight parameters z_j $(j = 1, ..., m)$ of SHG-Health were also set uniformly. The parameters α_l and α_u that control the influence from Y for the labeled and the unlabeled data respectively, were set to 0.01 for α_l as in [99] and tuned based on the grid $\{0.9, 0.99, 0.999, 0.9999, 0.99999\}$ for both SHG-Health and GGSSL methods.

TABLE 5.3: Evaluation on binary prediction (avg±std%)

	Precision	Recall/Sensitivity	F-Score
SHG-Health	96.24 ± 1.60	**43.93** ± 1.11	**60.32** ± 1.23
SHG-Health-Chi2	**99.33** ± 0.36	43.02 ± 1.40	60.02 ± 1.41
SHG-Health-Gaus	96.99 ± 1.32	43.69 ± 1.14	60.23 ± 0.86
SVM	89.00 ± 10.19	0.49 ± 0.36	0.98 ± 0.71
KNN	37.52 ± 1.48	25.62 ± 1.30	30.45 ± 1.36
GNetMine	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
GGSSL	5.21 ± 0.02	100.00 ± 0.00	9.91 ± 0.03

All of the experiments were conducted using 5-fold stratified cross validation. The performance on the testing set was obtained by averaging over 5-fold results. All the experiments were run on an Intel(R) Core(TM) CPU@3.40GHz workstation with 16GB physical memory.

5.4.2 Result Analysis

In this section we report and analyze the experimental results. The best $\lambda_{ij}, \forall i, j \in \{1, ..., m\}$ for GNetMine was 0.2 in all of our experiments, the same as reported in [68]. The best α_u for SHG-Health was 0.99, while GGSSL tended to bias toward one disease class with $\alpha_u = 0.9999$ and completely toward the unknown class with α_u less than that.

Identifying High-risk Cases

In this first stage of the two-stage evaluation (Section 5.4.1), we compared algorithms based on their abilities to identify high-risk cases regardless of what disease category they belonged to. All of the cases from different high-risk disease classes were regarded as belonging to one class, i.e., the high-risk class. This binary setting evaluated how well an algorithm was able to pick up high-risk cases in general.

Table 5.3, above, shows that our algorithms achieved the best overall performance at 99.33% precision (SHG-Health-Chi2), 43.93% recall (SHG-Health), and 60.32% F score (SHG-Health). GGSSL had 100.00% recall but extremely low precision at 5.21%, which indicates that it leaned towards predicting most cases as high-risk. On the other hand, GNetMine was completely biased toward the "unknown" class and thus had zero precision, recall, and F scores. The fact that recall scores for

TABLE 5.4: Evaluation on disease class prediction (avg±std%)

	Macro-Precision	Macro-Recall
SHG-Health	89.14 ± 0.56	89.62 ± 0.38
SHG-Health-Chi2	**90.58** ± 0.19	**90.73** ± 0.15
SHG-Health-Gaus	89.55 ± 0.56	90.30 ± 0.41
KNN	21.12 ± 1.49	59.92 ± 2.50
SVM	52.50 ± 39.41	63.33 ± 30.55
GNetMine	-	-
GGSSL	0.11 ± 0	9.09 ± 0

all methods were less than 50% also shows that capturing positive cases from large and noisy unlabeled cases was difficult. Our proposed SHG-Health could be seen as a conservative model, which is desirable for a preventive care system because the cost of false alarms is high [86, 63].

Classifying into Correct Disease Categories

In the second stage, we further evaluate the algorithms' *conditional* performance on multi-class classification. This is to evaluate the accuracy of disease-level predictions, given that the cases had been predicted into one of the disease classes. The macro-averaging measures were used to evaluate how correct these predictions were at the disease category level. Note that any "alive" case incorrectly predicted as one of the disease classes was counted as an incorrect prediction in this calculation.

Table 5.4 shows that our SHG-Health, especially SHG-Health-Chi2, outperformed all the other algorithms, achieving 90.58% macro-precision and 90.73% macro-recall. Overall, our algorithm was able to classify high-risk individuals into a correct disease category quite accurately.

Top Scored Test Items

Based on the COD labels available for the record type nodes, SHG-Health also computed scores as soft labels for other types of nodes, such as the *Physical Test* nodes. Within-class scores of a node type can reveal the relative importance of those nodes in the class.

Top-5 scored *Physical Test* items for 10 disease classes are listed in Table 5.5. There were some interesting results identified by our clinical experts. For example, for lung related disease categories, namely *pneumonia*, *chronic lower respiratory diseases*, and *malignant neoplasm of bronchus and lung*, chest examination had the highest score. This bears the interpretation that participants with

TABLE 5.5: Top 5 scored test items for 10 disease classes

Top 5	1) Pneumonia, organism unspecified	2) Chronic lower respiratory diseases	3) Malignant neoplasm of bronchus and lung
1	Chest Exam	Chest Exam	Chest Exam
2	Albumin	Urinary casts	Alpha-Fetoprotein
3	Urinary casts	Albumin	Hemoglobin
4	Hemoglobin	Hemoglobin	Hematocrit
5	Blood urea nitrogen (BUN)	Alpha-Fetoprotein	Mean corpuscular volume

Top 5	4) Malignant neoplasm of liver and intrahepatic bile ducts	5) Malignant neoplasm of Stomach	6) Malignant neoplasm of colon
1	Alpha-Fetoprotein	Albumin	Albumin
2	Aspartate Aminotransferase (GOT)	Hemoglobin	Alpha-Fetoprotein
3	Glutamic-Pyruvic Transaminase (GPT)	Mean corpuscular volume	Chest
4	Platelet count	Platelet count	Hemoglobin
5	Glucose urine test	Hematocrit	Creatinine blood test

Top 5	7) Chronic ischaemic heart disease	8) Malignant neoplasm of pancreas	9) Malignant neoplasm of gallbladder
1	Creatinine blood test	Glucose urine test	Alpha-Fetoprotein
2	Blood urea nitrogen (BUN)	Albumin	Aspartate Aminotransferase (GOT)
3	Glucose urine test	Mean corpuscular volume	Chest
4	Protein in Urine (Proteinuria)	Alpha-Fetoprotein	Mean corpuscular volume
5	Hemoglobin	Pus Cell in Urine	Red blood cell count

Top 5	10) Parkinson's disease
1	Urinary casts
2	Chest
3	Hemoglobin
4	Albumin
	Red blood cell count

these diseases commonly have abnormal chest examination results. Another obvious example can be found in the *malignant neoplasm of liver and intrahepatic bile ducts* class. Top-1 ranked item Alpha-Fetoprotein is a commonly used tumor marker for liver cancer. GOT and GPT are the enzymes concentrated in the liver, commonly used as key indicators for evaluating liver damage.

These results show that SHG-Health was able to identify important examination items for disease classes. By modeling features (i.e., examination items) as different types of nodes on a graph, the computing of soft labels for these nodes is actually a mechanism of feature weighting. It is the connection to these highly scored features of a class that determines the class label of a *Record* node in the graph.

Stability on Synthetic Data

We further compared the stability of the algorithms using two groups of synthetic data generated based on the distribution of the real data. The first group was the balanced datasets with increasing

size per class in the range of $\{100, 300, 500, 1000\}$. The second group contained the datasets with increasing number of unlabeled cases in the scale of $\{1, 3, 5, 10, 15, 20\}$ times of the size of a labeled class. For the details on how these datasets were generated, please refer to Section 5.4. We evaluated algorithms in terms of their ability to identify high-risk cases (Task 1) and their ability to classify a high-risk case into the correct disease class (Task 2). The same measures as in the real dataset case were used. It suffices to report the F scores for Task 1 and macro-averaging scores for Task 2.

Figures 5.4 to 5.8 below show the results from the balanced synthetic datasets. It can be seen from the F scores that all the algorithms performed stably in Task 1 except that 1NN had lower scores when class sizes were small. Our method was comparable to SVM and they had the highest performance in Task 1 (Figure 5.4 to Figure 5.6). However, our algorithm achieved significantly better macro-precision and macro-recall scores than the other algorithms in Task 2 (Figure 5.7 to Figure 5.8). While our SHG-Health stably maintained 70% macro-precision, others fluctuated below 50%. A similar phenomenon can be observed in the case of macro-recall.

We expected the performance would drop with the increasing scales of unlabeled data as shown in Figure 5.9 to Figure 5.11. The F scores of Task 1 confirm this intuition in Figure 5.11. The 1NN, GNetMine, and GGSSL methods had a steeper descending gradient than SVM and our approach, as the size of the unlabeled cases increased from 1,000 to 20,000. However, when it came to predicting correct disease classes (Task 2), as shown in Figure 5.12 and Figure 5.13, our SHG-Health had the highest macro-precision and macro-recall scores on the synthetic datasets. Note that the performance of our method went up as the unlabeled sizes increased from 10,000 to 20,000. It could be that more unlabeled cases helped our method to differentiate between disease classes better. It is worth mentioning that SVM had the most stable performance in Task 2, slightly below 50% for both macro-averaging measures.

Time Analysis

To investigate the scalability of all the algorithms, we recorded the training time for experiments on the synthetic datasets (Chapter 5.4.2), except that the testing time of 1NN was recorded. Figure 5.14, below, compares the time performance of all the algorithms with increasing data sizes. It can be seen that our method was the most time-efficient method of all, with only 1.41 seconds of training

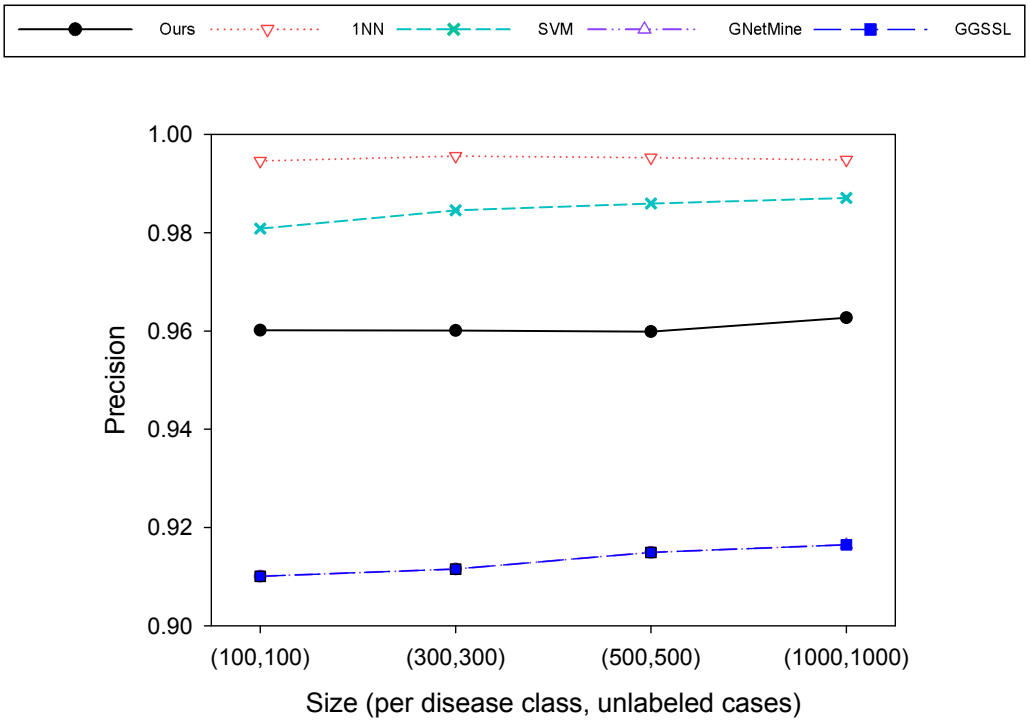

FIGURE 5.4: Task 1 results of the synthetic datasets with increasing unlabeled cases (Precision).

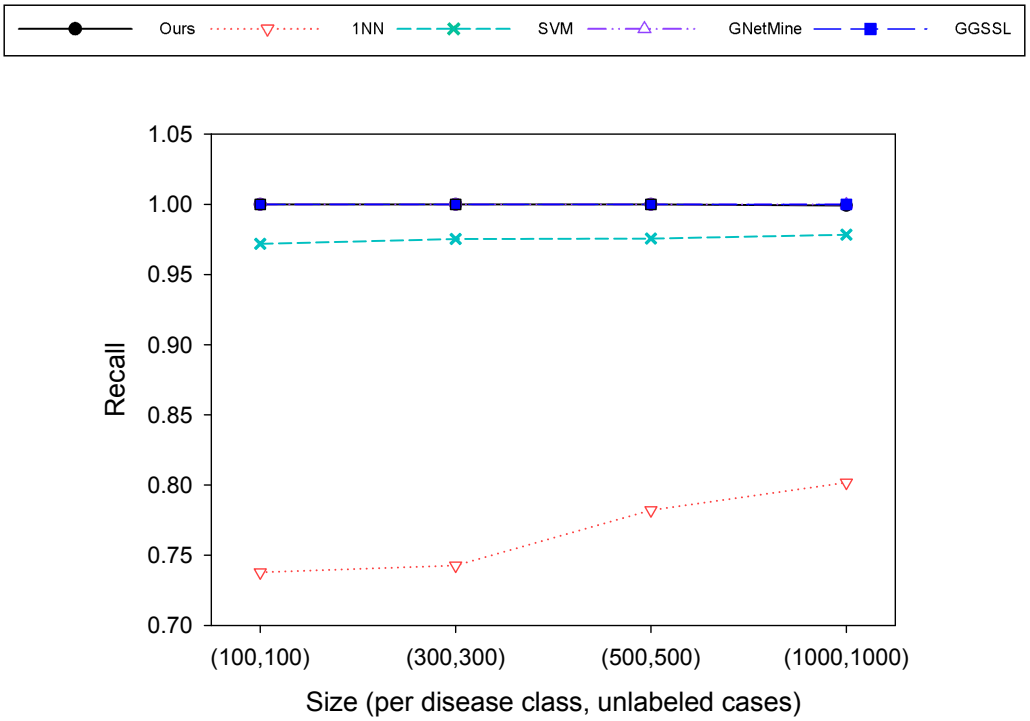

FIGURE 5.5: Task 1 results of the synthetic datasets with increasing unlabeled cases (Recall).

time at $(1000, 20000)$, i.e., 10×1000=10,000 disease cases plus 20,000 unlabeled cases. Note that the implementations we used for other methods are either standard implementations (SVM and 1NN) or from the author provided codes (GNetMine and GGSSL) and they might not be best tuned for

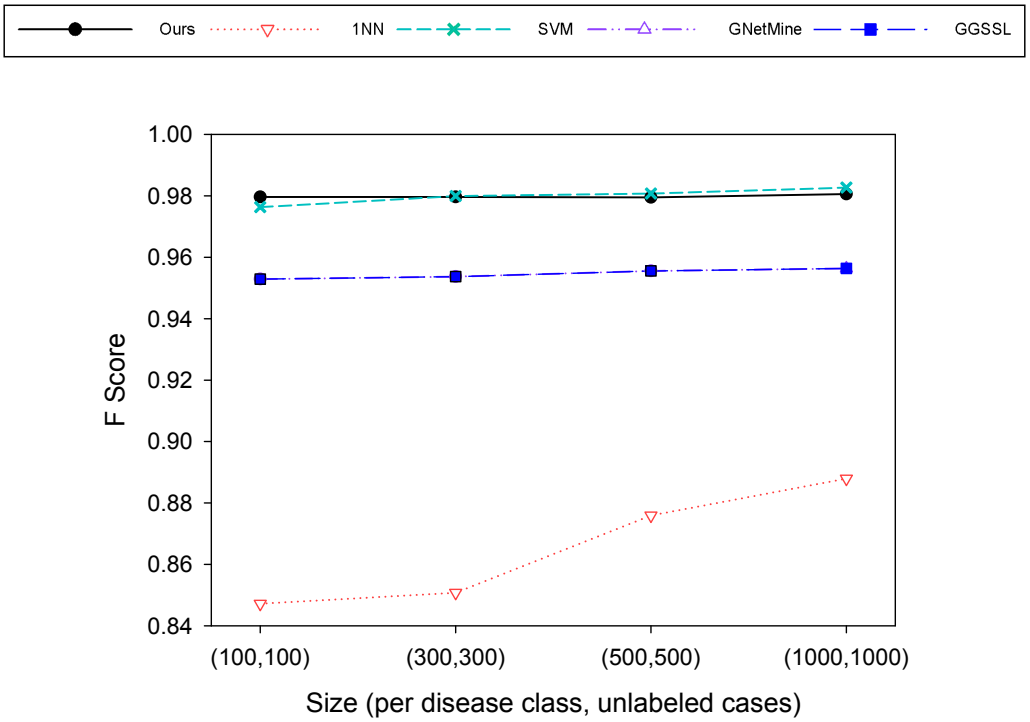

FIGURE 5.6: Task 1 results of the synthetic datasets with increasing unlabeled cases (F Score).

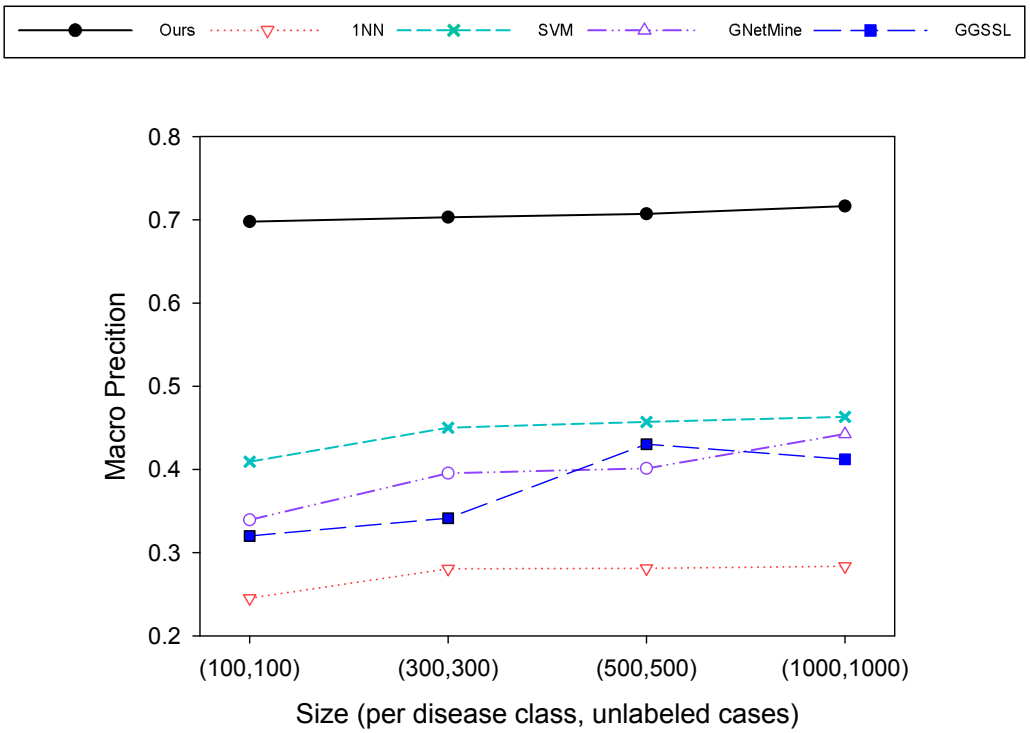

FIGURE 5.7: Task 2 results of the balanced synthetic datasets with increasing class sizes (Macro-precision).

efficiency. The purpose of Figure 5.14 is to show that our approach also enjoys desirable efficiency given its superior effectiveness as demonstrated in previous sections.

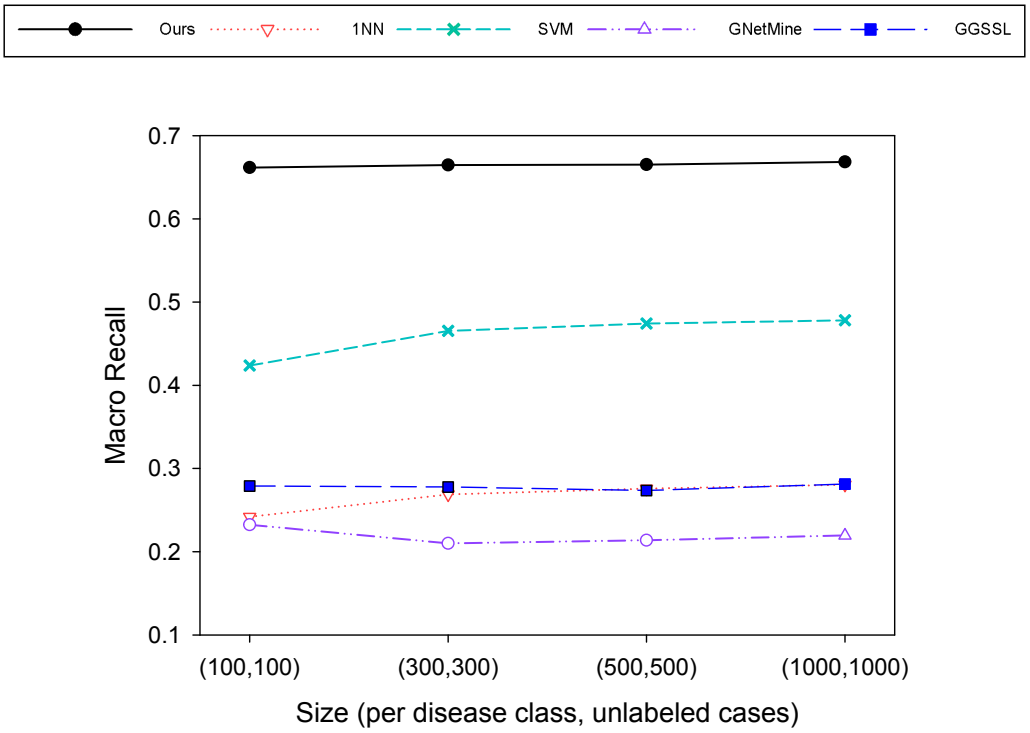

FIGURE 5.8: Task 2 results of the balanced synthetic datasets with increasing class sizes (Macro-recall).

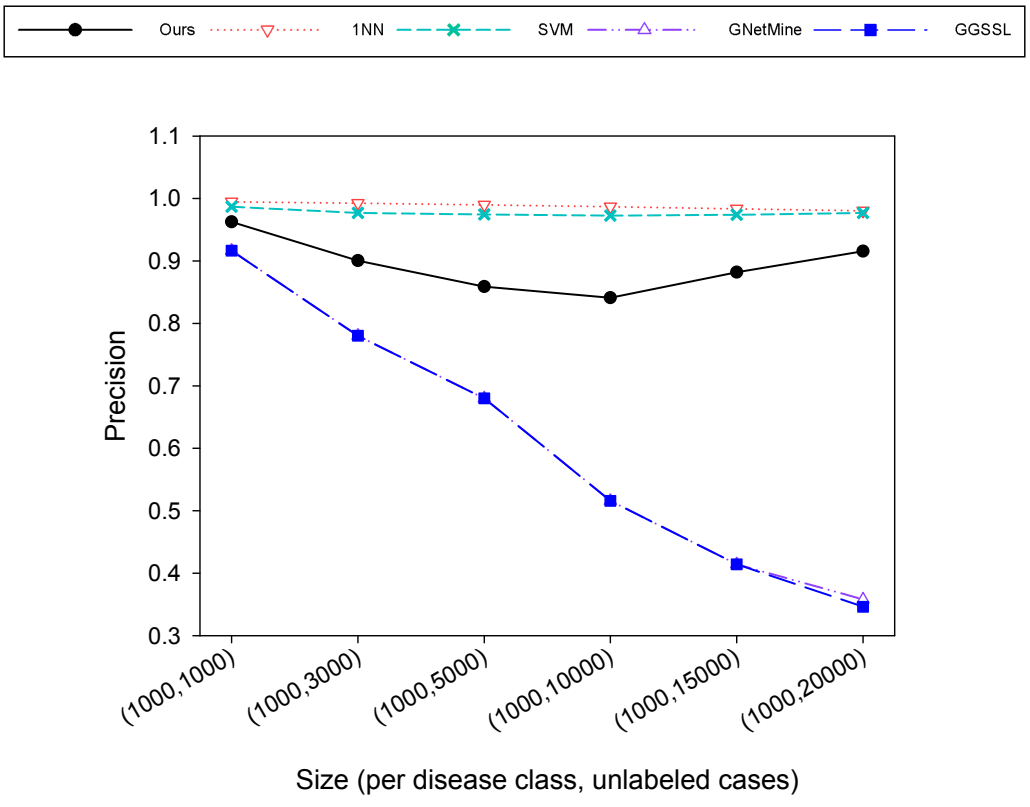

FIGURE 5.9: Task 1 results of the synthetic datasets with increasing unlabeled cases (Precision).

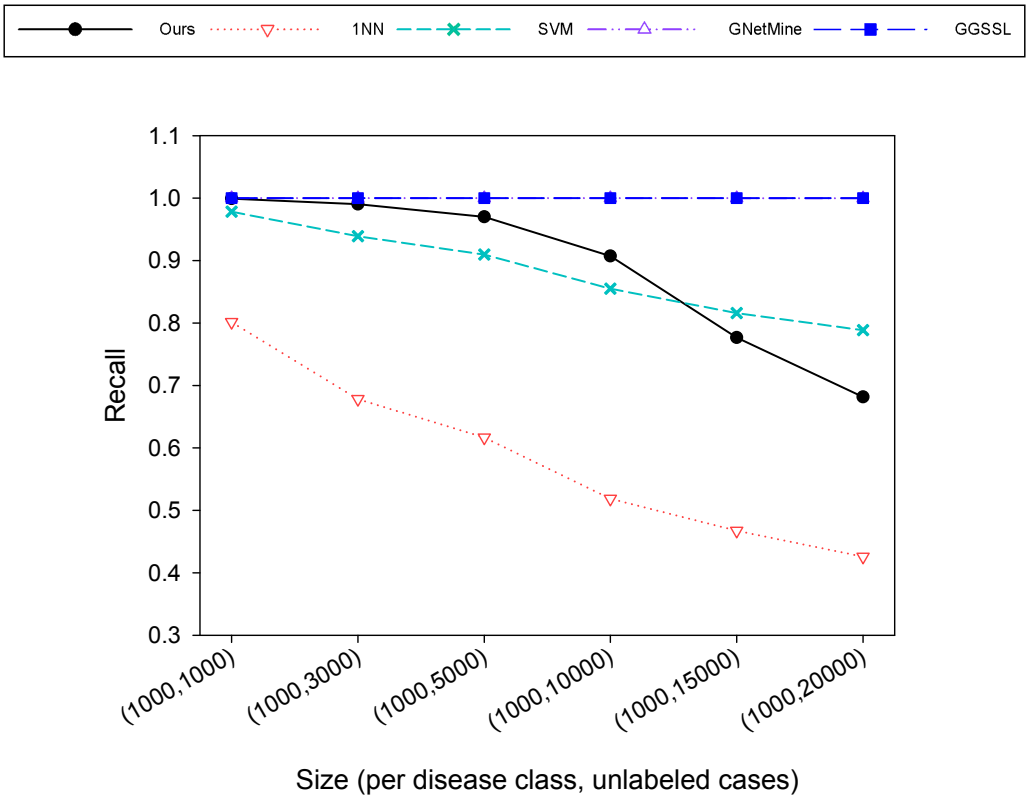

FIGURE 5.10: Task 1 results of the synthetic datasets with increasing unlabeled cases (Recall).

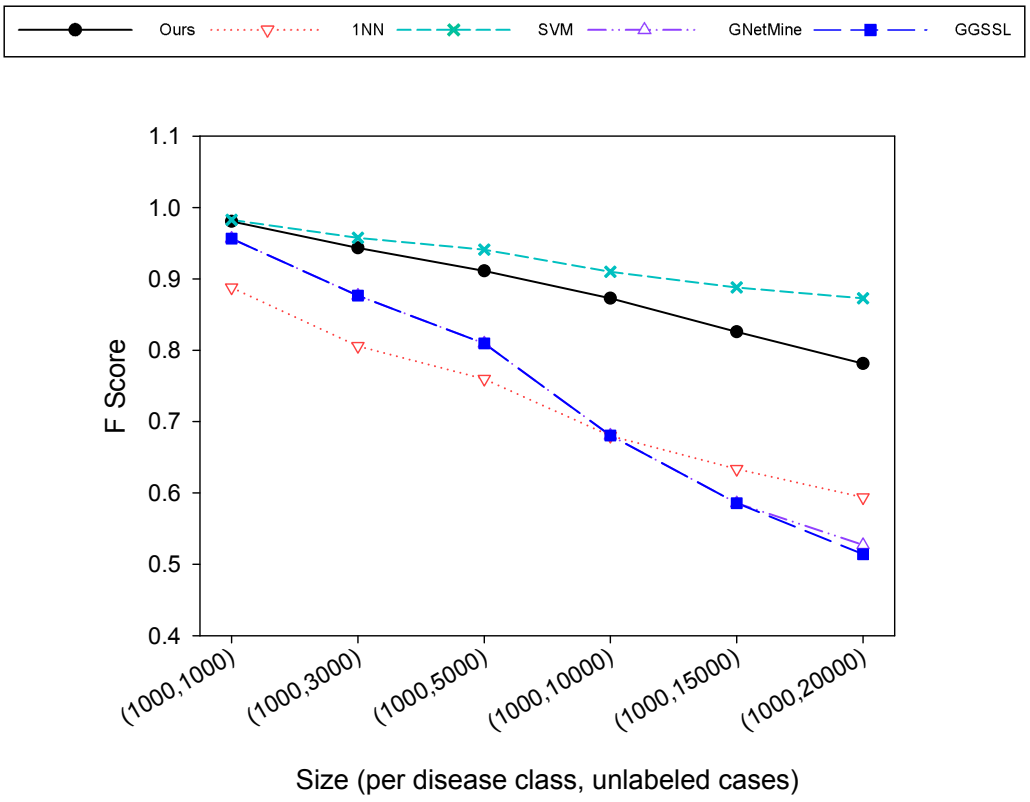

FIGURE 5.11: Task 1 results of the synthetic datasets with increasing unlabeled cases (F score).

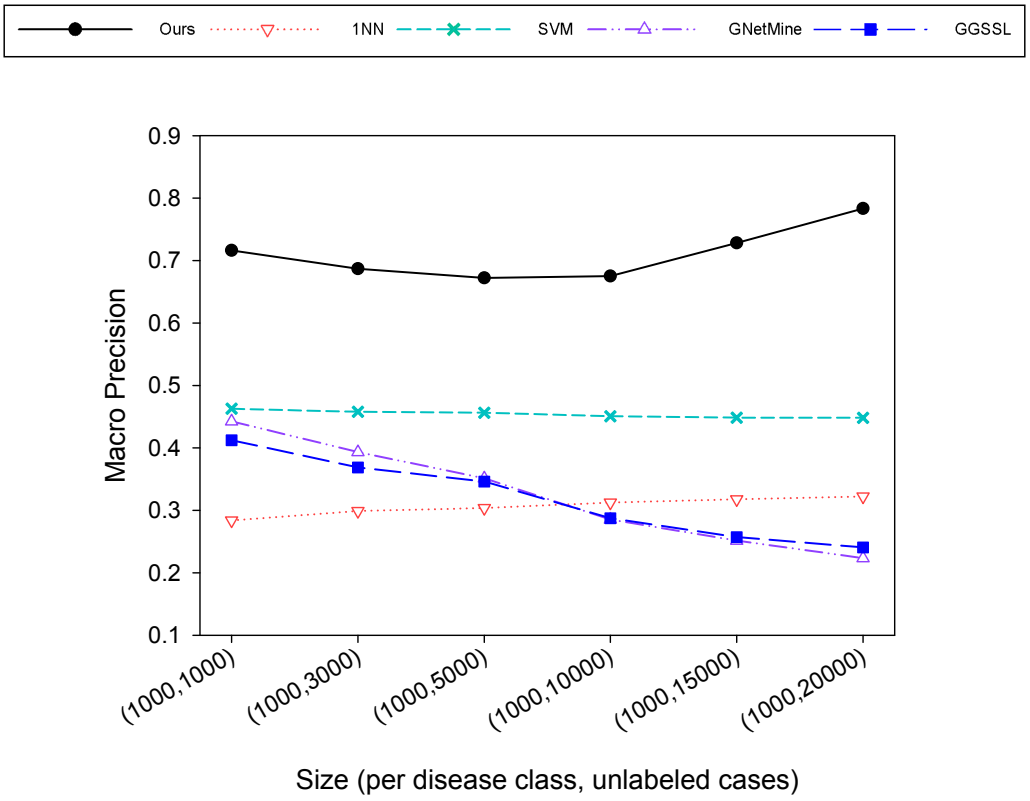

FIGURE 5.12: Task 2 results of the synthetic datasets with increasing unlabeled cases (Macro-precision).

5.4.3 Discussion

The experimental results showed that SHG-Health performed the best amongst all the algorithms compared. Particularly, its ability 1) to identify high-risk cases and 2) to predict the correct disease category for high-risk cases has been demonstrated and verified on our real datasets (Chapter 1.2.1) and synthetic datasets (Section 5.4).

From the results of real datasets reported in Table 5.3 and Table 5.4, we could see that although GNetMine also utilizes a heterogeneous network structure for classification, it tended to be biased toward the noisy "unknown" class. The reason is that GNetMine does not have a mechanism to control the label influence at instance level, such as the α parameters (Section 5.3.3) in our method and in GGSSL, nor the ability for class discovery. On the other hand, although GGSSL had such a mechanism, a homogeneous graph construction misses out the type-specific information that can help the classification. As a result, GGSSL was completely biased toward the dominant disease class.

In the case of synthetic datasets, the performance for most methods dropped in Task 2 (Figure

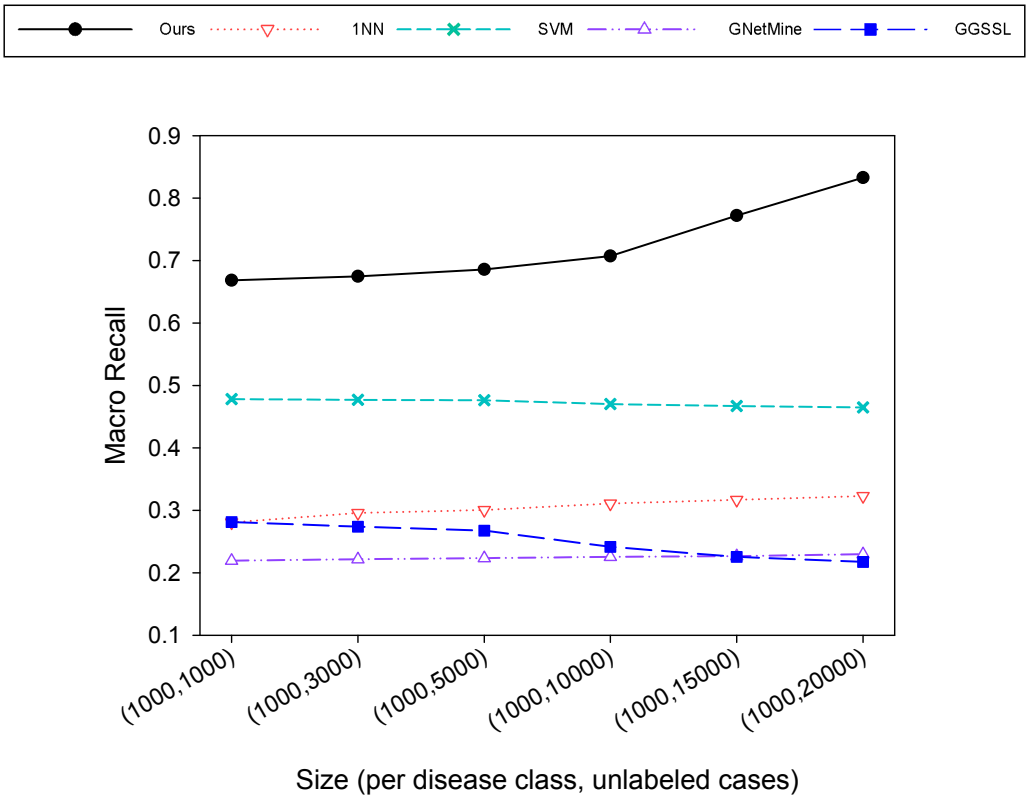

FIGURE 5.13: Task 2 results of the synthetic datasets with increasing unlabeled cases (Macro-recall).

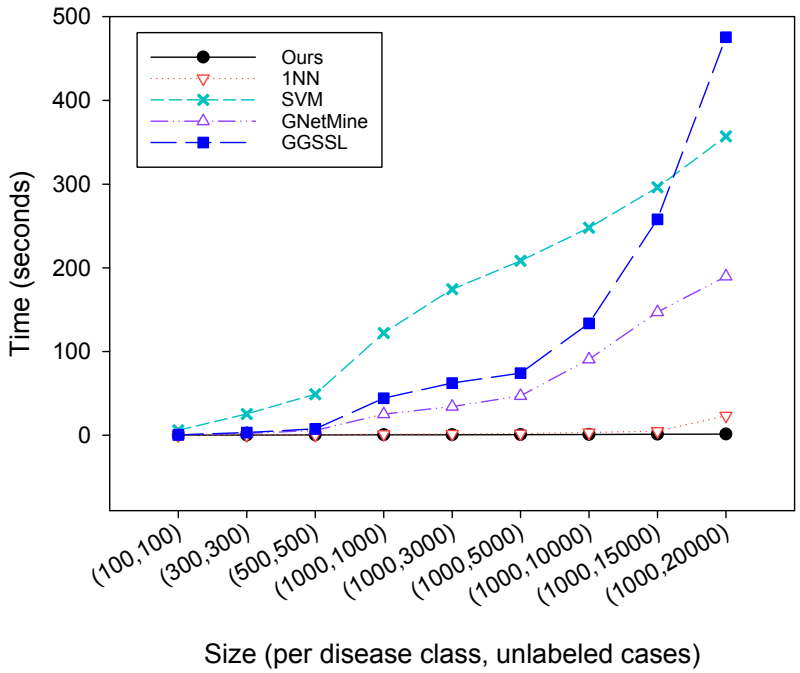

FIGURE 5.14: Computational time analysis on the synthetic datasets for the algorithms compared.

5.7 to Figure 5.8 and Figure 5.12 to Figure 5.13). This can be explained by the information lost due to the feature independence assumption for generating the synthetic datasets. The exceptions were GNetMine and GGSSL, which showed more reasonable performance on synthetic datasets. It could be that they were better in handling independent features than correlated ones. The efficiency of our method has also been demonstrated in Figure 5.14 based on the same synthetic datasets.

The above analysis demonstrated that our proposed SHG-Health is superior because it has the ability to handle the heterogeneity and large unlabeled data from health examination data. Although SHG-Health is conservative in predicting cases into high-risk classes, we have shown that it was able to predict the correct disease classes with high scores in all evaluation measures. This is very desirable for the preventative type of Clinical Decision Support Systems (CDSSs). False positives are especially costly in preventive care, which could result in unnecessary anxiety, worry, and invasive diagnostic tests [86, 63]. In addition, it is believed that CDSSs are to *support* clinical professionals rather than to *replace* them. Therefore, a good system should be able to identify and draw attention to participants with high risks.

5.5 Summary

Mining health examination data is challenging especially due to its heterogeneity, intrinsic noise and particularly the large volume of unlabeled data. In this chapter, we introduced an effective and efficient graph-based semi-supervised algorithm namely SHG-Health to meet these challenges.

Our proposed graph-based classification approach on mining health examination records has a few significant advantages over existing approaches.

- Firstly, medical examination records are represented as a graph that associates all relevant cases together. This is especially useful for modeling abnormal health examination results that are often sparse.

- Secondly, multi-typed relationships of data items can be captured and naturally mapped into a heterogeneous graph. Particularly, the medical examination items are represented as different types of nodes on a graph, which enables our method to exploit the underlying heterogeneous

subgraph structure of individual classes to achieve higher performance.

- Thirdly, features can be weighted in their own type through a label propagation process on a heterogeneous graph. These in-class weighted features then contribute to the effective classification in an iterative convergence process.

Chapter 6

Drug Recommendations from Medical Literature

In the previous chapters, we focused on classification problems for health risk/score prediction based on real healthcare datasets. In this chapter, we turn to explore another useful data source, namely medical literature. In particular, we look at a drug recommendation problem based on the MEDLINE medical literature dataset. This research was conducted based on a collaboration with The Third Xiangya Hospital, Changsha, China.

6.1 Problems and Challenges

In the age of information explosion, big data has the potential to improve clinical decision-making and revolutionize research [76]. The past decade has witnessed a dramatic growth of electronic biomedical literature, which is generated and stored in digital form and provides a substantial amount of reliable data that forms a rich source for Evidence-Based Medicine (EBM) research. However, it is no longer possible to process all the information available manually in order to keep clinical knowledge up to date.

To cope with this challenge, on the one hand, methods have been developed to improve document retrieval from biomedical literature databases [78, 136, 54, 13]. On the other hand, text mining techniques have been developed for extracting useful information from biomedical literature databases

[32, 114, 72, 70, 59] such as the MEDLINE database [3], a bibliographic database of life science literature, maintained by the U.S. National Library of Medicine. A comprehensive review on literature-based discovery for novel therapeutic approaches can be found in [59].

In this chapter, we focus on a drug recommendation problem based on the MEDLINE database. In particular, we deal with an antihypertensive drug recommendation problem for chronic kidney disease patients. We approached the problem with a graph-based data mining algorithm. Specifically, our method constructs a heterogeneous graph by extracting objects of different types from the MEDLINE medical literature database. We adapted our previously developed graph-based ranking algorithm MedRank [20] to utilize link structure of a heterogeneous graph for drug recommendation.

The rest of the chapter has been organized as follows. Section 6.2 gives the background of the Chronic Kidney Disease and literature-based EBM. Section 6.3 describes our methodology, including data sources, graph construction, and the ranking mechanism on the extracted heterogeneous graph. In Section 6.4, we evaluated the system recommended results with expert opinions collected via a questionnaire. Section 6.5 concludes our work in this chapter.

6.2 Background

Chronic kidney disease (CKD) is a worldwide public health problem that affects millions of people from all racial and ethnic groups [77]. In China, for example, the morbidity of CKD was approximately 10.8%, and its prevalence was estimated to be 119.5 million [139]. Hypertension is a major risk factor for the progression of CKD. It was reported to be present in 36% to 84.1% of patients with CKD [119] and contributed to progression toward End-Stage Renal Disease (ESRD) and cardiovascular events [101, 118]. However, the rate of blood pressure (BP) control was extremely low. Only 11% of patients reached their target goals [31] and the rate was even lower in developing countries [115].

In the era of Evidence-Based Medicine (EBM), medical decision-making is often optimized by emphasizing the use of evidence from well-designed and well-conducted studies. Although EBM helps physicians make decisions for medical care, in clinical practice, it suffers from possible absolutisation of the evidence obtained from Randomized Controlled Trial (RCT) and overestimating the

generalisability of the evidence [142]. Random sampling has been a highly successful mainstream method of clinical medical research. However, its accuracy depends on ensuring randomness, which is costly and difficult. Moreover, evidence-based clinical decision support does not adequately address the complexity of care and can be soon out of date.

These limitations of EBM have motivated us to develop an antihypertensive drug recommendation data mining application for CKD patients with hypertension.

6.3 The Methodology

To approach the problem of evidence-based antihypertensive drug recommendation for patients with CKD, we constructed a heterogeneous graph from the MEDLINE database and adapted a graph-based ranking algorithm called MedRank [20] to rank objects on the graph. Particularly, we ranked the antihypertensive drugs specified by our medical experts. MedRank is a graph-based algorithm that ranks objects (such as journal articles, diseases, treatments, authors, journals, etc.) in terms of referential relationships expressed as a heterogeneous graph. It was originally proposed to utilize linkage information among data objects to rank influential treatments for a given disease based on a medical information network. In this study, MedRank has been adapted to rank antihypertensive drugs on a heterogeneous graph extracted from a medical literature database.

In this section, we first introduce the data sources of this study, namely the MEDLINE [3] medical literature database and the MeSH [4] ontology that MEDLINE uses to index articles. Then we describe the process of extracting a heterogeneous graph from the MEDLINE database. Finally, we introduce our graph-based ranking method that ranks the extracted heterogeneous graph for drug recommendation.

6.3.1 Data Sources

MEDLINE and MeSH

The biomedical literature database employed in this study is the MEDLINE database [3]. It consists of more than 23 million journal citations and abstracts. The database is maintained by the National

Library of Medicine (NLM) and is indexed by the NLM Medical Subject Headings (MeSH) [4].

MeSH is a hierarchically structured medical thesaurus and controlled semantic vocabulary that is a part of the larger Unified Medical Language System (UMLS) thesaurus of the NLM. There are 27,883 descriptors more than 232,000 Supplementary Concept Records (SCRs) that provide formal and explicit specifications of the present biomedical concepts. The descriptors are arranged both alphabetically and hierarchically as a tree structure.

MEDLINE corpus can be acquired in XML format. Each citation contains the bibliographical information of an article, such as the article ID (PMID), article title, author list, journal title, venue, publication type, and indexed MeSH terms. MEDLINE is used as a surrogate for full-text articles. Permission to access the data in this study was acquired by The Third Xiangya Hospital in China in June 2015.

6.3.2 MEDLINE Article Selection Criteria

The MEDLINE articles considered in this study were studies on human hypertension combined with CKD. We used the MeSH headings indexed in MEDLINE to select these articles. The following four categories of selection criteria must be satisfied for an article to be qualified.

Human Criteria

- The indexed MeSH terms of an article must include “Humans”.

Disease Criteria

- The indexed MeSH terms of an article must include “Hypertension”.

- The indexed MeSH terms of an article must include either “Kidney Failure” or “Renal Insufficiency, Chronic”.

Drug Criteria

The antihypertensive drugs considered in this study belong to eight major categories of antihypertensive drugs, namely, diuretics, adrenergic $\alpha-$antagonists, adrenergic $\beta-$antagonists, angiotensin converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), calcium channel blockers (CCBs), vasodilator agents, and ganglionic blockers.

Articles on these antihypertensive drugs can be identified via the indexed MeSH terms. The MeSH terms for antihypertensive drugs that belong to the eight major categories include 95 heading terms and 121 supplementary concepts. To build a model for treatment recommendations, only the articles indexed by the above MeSH terms for antihypertensive drugs that were labeled with the qualifiers "therapeutic use", "drug therapy", or "administration & dosage" were extracted for modeling.

Publication Type Criteria

The publication type and publication date information of the articles were extracted for the purpose of article selection. The publication types under consideration include any of the following indexed MeSH terms: "meta-analysis", "randomized controlled trial", "pragmatic clinical trial", "twin study", "controlled clinical trial", "observational study", "comparative study", and "case report". We also extracted the publication year from the publication data information of the articles to indicate their recency.

6.3.3 Heterogeneous Graph Extraction

We constructed a weighted heterogeneous graph of five types of objects extracted from the MEDLINE corpus. These five object types are *Article*, *Author*, *Journal*, *Publication Type*, and *Antihypertensive Drug*. Every MEDLINE citation record that satisfied the "human" and "hypertension" selection criteria was modeled as an instance of the *Article* type object on the graph. The authors, publishing journal, publication type, and antihypertensive drugs associated with an article were modeled as instances of the *Author*, *Journal*, *Publication Type*, and *Antihypertensive Drug* types of objects respectively, with an edge to the Article object on the graph. The edge weight of each edge in the graph was calculated based on the following four factors:

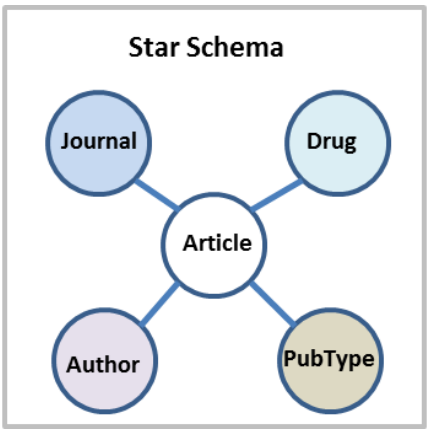

FIGURE 6.1: The schema extracted from the MEDLINE database.

- **Time factor (T)**: The publication year of an article was extracted as the time information about the article. If the article was published in the current year, the weight = 0; if it was published in the past 10 years, the weight = 1; otherwise, the weight = -1.

- **Publication type (P)**: Articles that report a meta-analysis or Randomized Controlled Trial (RCT), the weight = 1; otherwise, the weight = 0.

- **Design methods of clinical trials (D)**: If an article was indexed by random allocation, multi-center studies, double-blind method, and single-blind method, the weight = 1; otherwise, the weight = 0.

- **Institutions and funding sources (I)**: If the article was indexed by "Financing, Government", "Academies and Institutes", "Government", and "Organizations, Nonprofit", weight = 1; otherwise, weight = 0.

The final weight of a link is calculated by the function: 0.6+(T+P+D+I)/10.

6.3.4 Ranking on a Heterogeneous Graph

We adapted the MedRank algorithm [20] to rank the objects on the extracted heterogeneous graph. We then extracted a heterogeneous graph for our drug recommendation problem. As Figure 6.1 shows, the graph has a star-shaped schema, containing five object types, namely, the *Article*, *Author*, *Journal*, *Publication Type*, and *Antihypertensive Drug* object types, with *Article* as the center type.

In order to find the most influential drugs, all associated objects had to be considered. The reasoning behind the algorithm was as follows.

- A drug is "influential" if it is mentioned by many good articles, published with clinical trials that have positive results;
- An article is good if it is written by reputed authors and published in a good venue;
- A good venue is one that has a high impact on research; and
- An author is reputed if they have published many good articles.

At first glance, these definitions of "goodness" may seem to be circular. However, the algorithm is able to identify the listed above altogether by finding highly connected multi-objects on the graph.

The algorithm computes the ranks of objects of the same type based on the link relationship amongst objects of different types on the heterogeneous graph. It iteratively updates the ranks until convergence occurs using the following updating function Eq. (6.1) below:

$$R_{X_1}^{\langle h+1\rangle} \leftarrow (\alpha(\prod_{t=1}^{n-1} W_{X_tC} D_{CX_{t+1}}^{-1} W_{CX_{t+1}}) W_{X_nC} D_{CX_1}^{-1} W_{CX_1} + (1-\alpha)U/|X_1|) R_{X_1}^{\langle h\rangle} \tag{6.1}$$

where type index $t \in \{1, ..., n-1\}$ and n is a positive integer greater than 1 indicating the total number of types; X_t denotes the t^{th} object type and X_1 is the target type, i.e., the antihypertensive drug in our case; R_{X_1} is a vector for the rank of X_1 type objects; C is the center type, i.e., the *Article* in our case; U is an $|X_1| \times |X_1|$ unit matrix and $|X_1|$ denotes the total number of objects of type X_1; α is known as a damping factor that determines the ratio of the impact from the previous iteration and that from the global rank which is calculated uniformly as $U/|X_1|$; W_{AB} is a weighted adjacency matrix of type A and type B objects, storing the weighted links between them; D_{AB}^{-1} is a diagonal matrix with diagonal value equivalent to the row sum of W_{AB} for the purpose of row-normalization.

Mathematically, the algorithm can be seen as a random walk on a heterogeneous graph. The random walker decides which neighboring node to visit next based on the probabilities expressed as the link weights. Also the random walker might restart a new walk based on the probabilities controlled

Algorithm 4 Drug Recommendation Algorithm

Input: a set D of medical articles and a set of selecting criteria S.
Output: a vector $R_{X_1}^{\langle * \rangle}$ of the stationary ranks of the target type X_1 objects.

```
D_s ← select articles from D based on the selection criteria S.
W ← graph construction from D_s (Section 6.3.3).
Initialize R_{X_1} uniformly for type X_1 objects.
h = 1
repeat
    Update R_{X_1}^{⟨h⟩} by:
    R_{X_1}^{⟨h+1⟩} ← (α(∏_{t=1}^{n-1} W_{X_t C} D_{CX_{t+1}}^{-1} W_{CX_{t+1}}) W_{X_n C} D_{CX_1}^{-1} W_{CX_1} + (1 − α)U/|X_1|) R_{X_1}^{⟨h⟩}   (6.1)
    h = h + 1
until convergence return R_{X_1}^{⟨*⟩}
```

by the damping factors. The algorithm updates the probabilities of the links in each iteration. It terminates when it reaches stable probabilities, i.e., a stationary vector $R_{X_1}^{\langle * \rangle}$. This type of algorithms is able to discover highly connected regions of nodes in a graph. An object will be ranked high if its connected objects are also ranked high, which realizes the idea that a good article is written by good authors, published by a reputed venue, etc. The output of the algorithm is the ranking of the *Antihypertensive Drug* objects in its own type. The overview of our drug recommendation algorithm is summarized in Algorithm 4.

6.4 Experiments and Evaluation

A total of 4,304 papers, 13,325 authors, and 931 journals were extracted from the MEDLINE dataset based on the selection criteria described in the previous section for the experiment. These papers were published between 1965 and 2015 and included 420 comparative studies and 204 RCTs.

6.4.1 Settings

To compare the ranking differences in antihypertensive drugs over time, we extracted sub-graphs based on the ranges of publication year. The ranges ($\leq 1977, \leq 1980, \leq 1984, \leq 1988, \leq 1993, \leq 1998, \leq 2003, \leq 2013$) were selected according to the years indicated by the authoritative hypertension guidelines from the Joint National Committee (JNC) [17, 26, 48, 10, 27, 100]. The statistics of the articles in the corresponding JNC indicated years is shown in Table 6.1.

6.4.2 Questionnaire Design

A web-based questionnaire was designed for validating the ranked antihypertensive drug list by the proposed model. A survey was conducted according to the Declaration of Helsinki and was approved by the Medical Ethics Committee of the 3rd Xiangya Hospital (no. 2015-S169). The questionnaire was anonymous and the confidentiality of the demographic and personal data of each physician included in the survey was carefully preserved and strictly protected during each phase of the study.

Main Purpose

The main body of the questionnaire consisted of 22 single-choice or multiple-choice questions. The main purposes of the survey questionnaire were as follows:

- to collect the basic information of physicians, such as location, level of hospital, department, whether outpatient service was provided in the previous three months, position title, age, gender, and whether the hypertension guidelines were used (Questions 01 to 09);

- to obtain the physicians' preferences for antihypertensive prescriptions for hypertensive patients complicated with either CKD or diabetes (Questions 10 to 12 and 15 to 17). Physicians were asked to choose their first, second, and third preferences for the type and the name of antihypertensive drugs from a pre-specified list. Information regarding antihypertensive drugs is largely based on the MeSH terms; and

TABLE 6.1: Statistics over the years.

	Papers	Authors	Journals
1977	0	0	0
1980	0	0	0
1984	21	57	19
1988	165	553	101
1993	601	1,724	217
1997	1,017	3,044	313
2003	2,298	6,413	556
2013	5,363	17,399	1,144
2015	5,424	17,632	1,152

- to investigate physicians' preferences for optimal BP targets and their strategy of antihypertensive drug combinations when managing hypertensive patients with either CKD or diabetes (Questions 13 to 14 and 18 to 22).

Another aim of the survey was to analyze the concordance among the physicians included in the survey and to stratify the results according to area, age, gender, department, hospital level, and position of the physician (resident, attending, associate chief, or chief).

Sample Size

The following formula was used for calculating the sample size of physicians enrolled in the study:

$$N = 2 * z/(d2 * p * q)$$

where z is the risk of Type I error, d is the absolute precision, p is the expected prevalence, and q=1-p. Using a 95$\%$ confidence interval (95$\%$ CI), 80$\%$ power, an estimated precision of 5$\%$, and assuming the concordance rate between the system and physician rankings to be 65.8$\%$ according to [20], the estimated sample size was 168.

Survey Distribution

The survey was sent to members of the China Hypertension Alliance by email or WeChat Version 4.0 (WeChat, Tencent) using a snowball-sampling strategy. Recruitment of physicians occurred between October 2015 and November 2015. The participants involved in this survey included both general practitioners and specialized physicians operating in different clinical settings (outpatient clinics and/or in-hospital divisions), geographical locations (both north and south of China), and different clinical activities so that we could obtain a representative sample of physicians with experience in patients with hypertension in China. The participants did not receive any remuneration for taking part in the survey.

6.4.3 Evaluation Measures

All analyses were performed using IBM SPSS Version 22.0 (SPSS, Inc., Chicago, IL). The present analysis was limited to the answers provided in the main body of the survey questionnaire. Continuous variables were presented as the means±standard deviations (SDs) or as medians and interquartile ranges (IQRs) when the data exhibited a skewed distribution. Analysis of variance (ANOVA) or a *post hoc* analysis of the least significant difference (LSD) was used to compare the means of normally distributed continuous variables; otherwise, the Kruskal-Wallis test was used. Categorical variables were expressed as proportions and tested using the chi-square test. A P-value of 0.05 (two-tailed) was established as the threshold for statistical significance.

To compare the consistency between the preferences of physicians and the ranking results, the average overlap (AO)[38] measure was used. The AO is a measure of the concordance between two ordered lists that may not have equal lengths. It computes the weighted average of the overlap of the first d elements in both lists, with the weight decreasing as d increases. In other words, the overlaps that occur at the top of the list are counted as more important than are those at the end. The AO score is in the range of [0,1] inclusive, where 0 indicates no items shared by the two lists and 1 indicates that the two lists are identical. The formula is shown in Eq. (6.2).

$$AO(S, T, K) = \frac{1}{K} \sum_{d=1}^{K} \frac{|S_{:d} \cap T_{:d}|}{d} \tag{6.2}$$

where K is the top-K ranked items of interest; S and T are two ranking lists; $S_{:d}$ denotes the set of elements in S from the first up to the d^{th} position; $T_{:d}$ is defined in the same way as $S_{:d}$.

6.4.4 Results

According to the recommendations of our model, the top ranked categories of first-line antihypertensive drugs are ordered as follows: ARB, ACEI, CCB, diuretics, adrenergic -antagonists, vasodilator agents, ganglionic blockers, and adrenergic -antagonists. The top 10 ranked individual antihypertensive drugs for patients with CKD are reported in Table 6.2 with Irbesartan at the top. Although only six drugs that belong to RAS inhibitors in the table were consistent with the latest guidelines, the other

TABLE 6.2: Top 10 ranked hypertensive drugs.

Top K	Drug	Ranking	# of supported studies
1	Irbesartan	0.151	26
2	Ramipril	0.138	11
3	Benazepril	0.087	6
4	Losartan	0.08	17
5	Captopril	0.061	27
6	Amlodipine	0.058	12
7	Enalapril	0.049	20
8	Furosemide	0.048	39
9	Nitrendipine	0.041	3
10	Atenolol	0.041	10

kinds of drugs (adrenergic β-antagonists, diuretics, and CCB) were also used as first-line antihypertensive agents when there were contraindications to RAS inhibitors or used as drug combinations in CKD patients.

Note that the numbers in the ranking column in Table 6.2 are the limiting probabilities, so that the sum of ranks for all drug object in the extracted graph is equal to 1. As a result, the larger the number of antihypertensive drugs considered in the model, the smaller the ranking score for individual drugs.

The number of citations that support the antihypertensive drugs Irbesartan, Ramipril, and Benazepril were found to be 26, 11, and 6, respectively. Although the citation number contributed to the ranking, it was not a decisive factor.

Table 6.3 illustrates the top 5 ranked results of four year ranges. This can be compared against Table 6.4 where the recommendations from the hypertension guidelines of the Joint National Committee (JNC) [17, 26, 48, 10, 27, 100] are given. The top recommended antihypertensive drug by our method calculated from studies performed before 1980 was Furosemide, a type of loop diuretic, which was matched the recommendations by JNC 1 (1976) and JNC 2 (1980). The introduction of Adrenergic-inhibiting agents and ACEI in the JNC 3 (1984) guildline could explain the rising uses of these two types of drug, which was also reflected in the top 2 to 4 drugs by our method.

According to our model, loop diuretics became less important over time, and the importance of ARBs increased rapidly. In 1993, ACEIs became the top antihypertensive drugs in our model. But by 2013, ARBs became more influential than ACEIs and became the leading drugs. However, the JNC 5 [48], still emphasized the importance of loop diuretics in 1993. ACEIs were recommended by the

TABLE 6.3: Top 5 ranked hypertensive drugs (name,category) in year ranges.

Year Range	Top K	Drug Name
<1980	1	Furosemide, loop diuretic
	2	Captopril, ACEI
	3	Prazosin, adrenergic-inhibiting agents
	4	Labetalol, adrenergic-inhibiting agents
	5	Minoxidil, vasodilators
<1993	1	Captopril, ACEI
	2	Nitrendipine, CCB
	3	Furosemide, loop diuretics
	4	Enalapril, ACEI
	5	Minoxidil, vasodilators
<2003	1	Captopril, ACEI
	2	Nitrendipine, CCB
	3	Losartan, ARB
	4	Enalapril, ACEI
	5	Ramipril, ACEI
<2015	1	Irbesartan, ARB
	2	Ramipril, ACEI
	3	Benazepril, ACEI
	4	Losartan, ARB
	5	Captopril, ACEI

JNC 6 as the first-line drug for patients with CKD until 1998 [10]. Figure 6.2 shows the changes of ranking order of antihypertensive drug categories over the years computed by our model.

Overall, 377 physicians (174 males, mean age 35.65±9.89 years) were included in this present study, and all completed the survey, including 49.34% who had regularly treated outpatients in the previous three months. The vast majority of clinicians were cardiologists (206, 54.64%), followed by specialists in general medicine (23, 6.1%), nephrology (23, 6.1%), and surgery (20, 5.31%). The distribution of physicians nationally was substantially homogeneous. A total of 59.63% of physicians were active in Southern China, and 40.37% were active in Northern China. Most (76.92%) worked at tertiary hospitals. According to this web survey, 44.30% (167/377) of physicians preferred to prescribe ACEIs as their first choice for patients with CKD, whereas only 28.12% (106/377) chose CCBs, and 25.73% (97/377) chose ARBs. Generally, the concordance between system ranking and physician ranking was not high (AO, 0.53±0.20). Interestingly, as the position of a drug increased on the list, the concordance increased, as shown in Figure 6.3. However, the AO scores did not significantly differ according to area, age, gender, department, and hospital level.

TABLE 6.4: Guidelines from the Joint National Committee (JNC)

Version	Year	Quotes
JNC 1	1976	'Furosemide is not usually a step 1 drug but is especially useful in renal insufficiency or resistance cases'.
JNC 2	1980	No change
JNC 3	1984	'loop diuretics are effective in chronic kidney failure'; 'Adrenergic-inhibiting agents, vasodilators, ACEIs are usually effective in patients with renal disease'
JNC 4	1988	'loop diuretics are effective in chronic kidney failure'; 'Recent studies, particularly in diabetic nephropathy, suggest that the use of ACE Inhibitors may possess specific advantages in decreasing proteinuria and slowing the progression of renal failure'.
JNC 5	1992	'Higher doses of loop diuretics may be needed for patients with renal impairment'; 'A loop diuretic or metolazone or indapamide is usually necessary or accomplishes substantial diuresis when the serum creatinine level has reached 221 mol/l or more'.
JNC 6	1997	'...patients with hypertension who have renal insufficiency should receive, unless contraindicated, an ACE Inhibitor (in most cases, along with a diuretic) to control hypertension and to slow progressive renal failure'.
JNC 7	2003	'The ACE inhibitors and ARBs have demonstrated favorable effects on the progression of diabetic and non-diabetic renal disease'.
JNC 8	2013	'In the population aged 18 years of age or older with CKD, initial (or add-on) antihypertensive treatment should include an ACEI or ARB to improve kidney outcomes. This applies to all CKD patients with hypertension regardless of race or diabetes status. (Moderate Recommendation Grade B)'.

6.5 Summary

In this chapter we demonstrated that drug recommendations based on medical literature can be computed based on the graph-based ranking algorithm that we proposed. Particularly, we verified our approach by evaluating our model recommended drug list against expert preference of antihypertensive drug prescription for CKD patients. The AO scores (Figure 6.3) showed that our recommendations were not only promising but had a high concordance with the preferences of the physicians with higher ranks. In addition we computed drug recommendations based on articles published in different year ranges and compared the results against the JNC guidelines. We also showed that our model recommendations matched several key changes of the JNC guidelines (Figure 6.2 and Table 6.4).

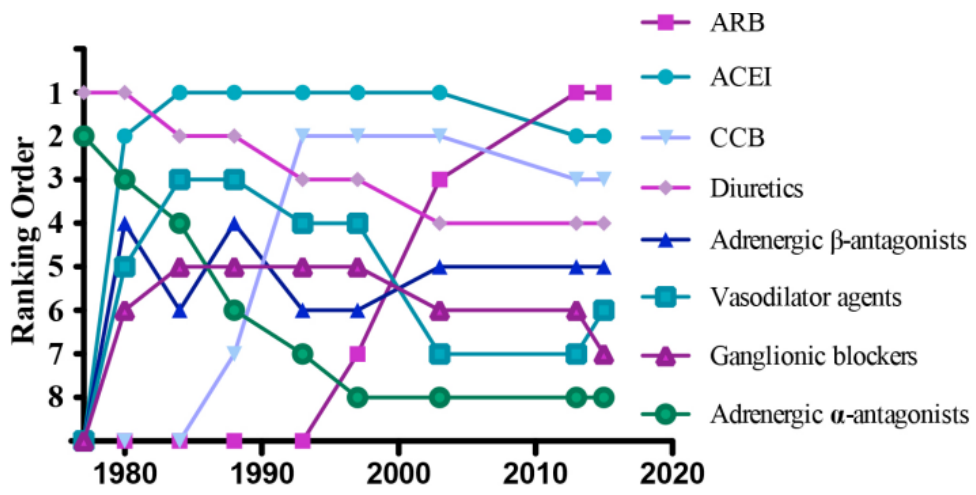

FIGURE 6.2: *Comparison of the Top 10 ranked antihypertensive drug categories over years.*

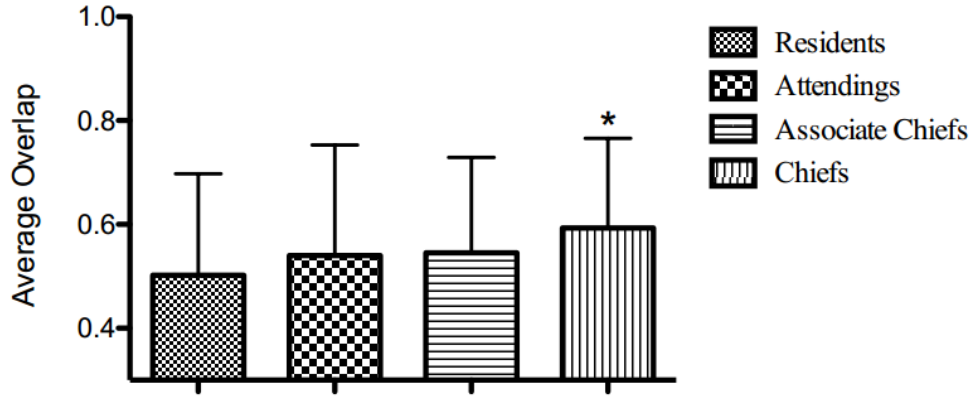

FIGURE 6.3: *Average Overlap scores in terms of physicians' position level.*

Our study opens up a new direction of evidence based decision support that combines the following features.

- The proposed literature-based data mining method is able to recommend antihypertensive drug categories prior to the release of hypertension guidelines.
- It recommends the most influential treatment for a given disease, which is specific to drug names, accounting for evidence level and time.

Chapter 7

Conclusion and Future Work

7.1 Conclusions

This thesis has been driven by the practical need for mining healthcare data from large and multiple data sources. Particularly, we focused on developing effective classification models for healthcare applications, tackling issues identified from real healthcare datasets. The main objectives have been successfully achieved. Specifically, we have accomplished three tasks:

- We identified problems and challenges in learning classification models for healthcare applications, namely feature extraction, label uncertainty, and large unlabeled data issues.
- To address these problems, we developed algorithms that were able to deal with those issues.
- We evaluated the proposed approaches with real-world health examination datasets and synthetic datasets.

In Chapter 1, we introduced the research motivation and background of healthcare data mining with an emphasis on health risk prediction, including the scope of the field and definition of Electronic Health Records (EHRs). In particular, we specified the main datasets that this research was based on in the background section, identified their characteristics, and gave a formal definition of General Health Examination (GHE) record sequence. Finally, we presented research challenges, goals, and the main contributions of this thesis.

In Chapter 2, we presented a brief literature review on the research topics that are related to our research. We divided the literature review into three sections. In the first section, we reviewed the traditional data analysis approaches in the medical literature. Particularly, we paid attention to the existing data analytics type work on health examination records as well as the studies on health scoring systems that are often used in critical care. In the second section, we reviewed the literature on mining healthcare data with an emphasis on classification problems, summarising existing studies on classification-based risk prediction. We also reviewed existing approaches to handling label uncertainty and large unlabeled data for healthcare applications. In the concluding section, we briefly reviewed the related work on text mining of medical literature.

In Chapter 3, we proposed a classification-based framework based on a large General Health Examination (GHE) dataset for overall health score prediction based on mortality labels. We called this predicted score "Personal Health Index" (PHI), which is defined as the complement probability of health risk. Our framework contained the following four stages:

- Pre-processing was performed to clean the raw data and transform the numerical, ordinal, and categorical features into binary representation.
- We then proposed two representation strategies for modeling record sequence were proposed. Particularly, we used time smoothing functions to merge record sequence into a feature vector. Our approach is capable of handling the time dimension of record sequence.
- With this we trained SVM classification models with different kernels and conducted feature selection.
- Finally, our framework was evaluated based on a real General Health Examination (GHE) dataset (Section 1.2.1). The experiments showed that the proposed PHI prediction framework could effectively predict PHI and identify high health risk participants.

In Chapter 4, we extended the PHI prediction framework described in Chapter 3 from an overall health score to a vector of k predicted scores in the range of $[0, 1]$ interval, reflecting personal health risks in k disease categories. This led us to use the main Cause of Death (COD) information as labels. However, due to the possible comorbidity, the COD label may not be 100% certain.

To tackle this classification problem with label uncertainty, our prediction framework contains four steps.

- Data pre-processing was conducted as in Chapter 3.
- We then extracted features based on a Chi-Square time smoothing function to handle the time aspect of record sequence.
- With this we proposed a soft-label learning approach that learns from target examples, related examples, and negative examples. In our case, these are the examples of the target disease, of the non-target diseases, and of the alive cases respectively.

We evaluated the proposed method on the same GHE dataset of 102,258 patients (Chapter 1.2.1). The experimental results demonstrated the effectiveness of our algorithm.

In Chapter 5, we tackled the large unlabeled data issue identify from our GHE dataset with linked COD labels. In fact only 7.4% of GHE participants have a COD label. The risk prediction problem was formulated as a multi-class classification problem with large unlabeled data. We proposed a graph-based semi-supervised algorithm called *SHG-Health* to solve the problem.

Two key steps are involved in our approach.

- A heterogeneous graph is constructed based on the GHE dataset. Specifically, four types of nodes were extracted, namely, *Record*, *Physical Test*, *Mental Test*, and *Profile*. In addition, the links between nodes were weighted according to the time of the examination took place.
- We proposed an efficient semi-supervised algorithm to iteratively learned the model based on the extracted heterogeneous graph.

Our approach has the following two advantages.

- It is capable of capturing the heterogeneous nature of GHE data.
- It has a mechanism to learn an additional "unknown" class based on the heterogeneous graph structure.

The proposed algorithm was evaluated using our GHE dataset as well as synthetic datasets. The experimental results showed that our method can effectively handle large unlabeled data and that it enjoyed a superior performance over baseline methods and other graph-based algorithms compared.

In Chapter 6, we turned to explore another data source, namely a biomedical literature database called MEDLINE. Specifically, we tackled the problem of antihypertensive drug recommendations for Chronic Kidney Disease (CKD) patients based on MEDLINE biomedical literature. Our approach contains two steps.

- We extracted from MEDLINE different types of objects via the indexed MeSH terms to construct a heterogeneous graph.
- We then adapted a graph-based algorithm to rank the drug objects on the graph for drug recommendation.

Our system was evaluated based on the expert opinions collected via questionnaire. The results showed that our system recommendations had a high concordance with expert preferences collected via a online questionnaire.

7.2 Future Directions

Overall, we believe that this research has provided a new direction of quantifying personal health through data mining techniques. We propose two possible directions for future investigation.

7.2.1 Data Fusion for Personalized Healthcare

Accurate, personalised, and cost-effective healthcare is the vision of the future. We believe that can be possible from cradle to grave, through trends and predictions derived from mining large real health datasets. Our research had its main focus on real health examination data and we also investigated the use of biomedical literature. Other information sources, such as online medical information, biomarks collected from mobile sensor devices (e.g., accelerometer and gyroscope on a smart phone recording the daily activities of a person), and in-hospital data, could be linked to obtain more comprehensive results. Skillful mining of such a fusion of data could well make the vision a reality.

7.2.2 Feature Extraction for Longitudinal Data

Health records stored according to date/time order form a time-series of records. Each record in the series may contain a large number of measurements. One big challenge of healthcare data mining is extracting important information from the high dimensional time-series for modeling. In our research, we had considered time smoothing functions to transform record sequence into a feature vector. We also used graph representation with time-weighted edges. However, how to extract features that are able to capture changes in high dimensional measurements over time is still an open problem. A systematic study on the effects of different representational approaches should be beneficial to this area of research.

7.3 Personal Reflection

My journey of completing this PhD study has been a journey full of the excitement of discovery, though sometimes with ups and downs. Developing classification models to improve clinical decision support is a research direction which I personally found especially significant because of its direct applications to healthcare.

I derived greatest satisfaction from knowing that we were working with real data of real people, knowing this would enable doctors to make more accurate, personalised, and efficient clinical judgements. It was not academic.

Perhaps the lowest point was the three-month period of data pre-processing at the beginning. It was full of uncertainties regarding whether the quality of the data would be good enough for sensible data mining, what assumptions would be reasonable for handling different features, and to what extent we might gain support from medical professionals. The tedious nature of the pre-processing often made me wonder if this was worthwhile at all. Now I can say, yes, by all means. The more one understands the data, the more they would be able to understand why when a model did not work and to come up with ideas that might work.

In this journey, the more I dug into the problem the more avenues of exploration opened up before me. At times I felt like I was only just scratching the surface. It is my humble hope that what we have done here will form a foundation from which others can go further.

References

[1] Health assessment for people aged 75 years and older. http://www.health.gov.au/internet/main/publishing.nsf/Content/mbsprimarycare_mbsitem_75andolder. Accessed: 2014-06-18.

[2] International classification of diseases. http://www.who.int/classifications/icd/en/. Accessed: 2015-05-05.

[3] MEDLINE Fact Sheet. https://www.nlm.nih.gov/pubs/factsheets/medline.html. Accessed: 2016-04-05.

[4] MeSH Fact Sheet. https://www.nlm.nih.gov/pubs/factsheets/mesh.html. Accessed: 2016-04-05.

[5] NuffieldHealthScore. http://info.nuffieldhealthscore.com/. Accessed: 2014-06-18.

[6] Predicting mortality of ICU patients. http://physionet.org/challenge/2012/. Accessed: 2014-02-18.

[7] Scikit learn: machine learning in Python. http://scikit-learn.org/stable/. Accessed: 2014-06-18.

[8] U.S. national health and nutrition examination survey. http://www.cdc.gov/nchs/nhanes.htm. Accessed: 2015-06-08.

[9] Your NHS health check guide. http://www.nhs.uk/Planners/NHSHealthCheck/Pages/Overview.aspx. Accessed: 2014-06-06.

[10] The sixth report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Archives of Internal Medicine*, 157(21):2413–2446, 1997.

[11] Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*, 7:326–337, 2002.

[12] C. C. Aggarwal. *Outlier Analysis*. Springer, New York, NY, 2013.

[13] P. Avillach, J.-C. Dufour, G. Diallo, F. Salvo, M. Joubert, F. Thiessard, F. Mougin, G. Trifiro, A. Fourrier-Reglat, A. Pariente, and M. Fieschi. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project. *Journal of the American Medical Informatics Association*, 20(3):446–452, 2013.

[14] S. Bandyopadhyay, B. Ganguli, and a. Chatterjee. A review of multivariate longitudinal data analysis. *Statistical methods in medical research*, 20(4):299–330, aug 2011.

[15] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1):56–68, 2011.

[16] Y. Baram, R. El-Yaniv, and K. Luz. Online Choice of Active Learning Algorithms. *Journal of Machine Learning Research*, 53(9):1689–1699, 2013.

[17] R. M. Carey, J. Cutler, W. Friedewald, N. Gant, S. Hulley, J. Iacono, M. Maxwell, D. McNellis, G. Payne, A. Shapiro, et al. The 1984 report of the joint national committee on detection, evaluation, and treatment of high blood pressure. *Archives of Internal Medicine*, 144(5):1045–1057, 1984.

[18] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[19] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. Mackenzie. a New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation. *J Chron Dis*, 40(5):373–383, 1987.

[20] L. Chen, X. Li, and J. Han. MedRank: Discovering Influential Medical Treatments from Literature by Information Network Analysis. In *Proc. 2013 Australasian Database Conf.(ADC'13)*, pages 3–12, 2013.

[21] L. Chen, X. Li, Q. Z. Sheng, W.-c. Peng, J. Bennett, H.-y. Hu, and N. Huang. Mining health examination records: A graph-based approach. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2423–2437, 2016.

[22] L. Chen, X. Li, S. Wang, H.-Y. Hu, N. Huang, Q. Z. Sheng, and M. Sharaf. Mining Personal Health Index from Annual Geriatric Medical Examinations. In *2014 IEEE International Conference on Data Mining*, pages 761–766, 2014.

[23] L. Chen, X. Li, Y. Yang, H. Kurniawati, Q. Z. Sheng, H.-Y. Hu, and N. Huang. Personal health indexing based on medical examinations: A data mining approach. *Decision Support Systems*, 81:54 – 65, 2016.

[24] Y. Chen, R. J. Carroll, E. R. M. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2):1–7, 2013.

[25] Y.-w. Chen and C.-j. Lin. Combining SVMs with Various Feature Selection Strategies. In *Feature Extraction*, number 1, chapter 12, pages 315–324. Springer Berlin Heidelberg, 2006.

[26] A. V. Chobanian, M. H. Alderman, V. DeQuattro, E. D. Frohlich, R. W. Gifford, M. N. Hill, N. M. Kaplan, H. G. Langford, M. A. Moore, W. A. Nickey, et al. The 1988 report of the joint national committee on detection, evaluation, and treatment of high blood pressure. *Archives of Internal Medicine*, 148(5):1023–1038, 1988.

[27] A. V. Chobanian, G. L. Bakris, H. R. Black, W. C. Cushman, L. A. Green, J. L. Izzo Jr, D. W. Jones, B. J. Materson, S. Oparil, J. T. Wright Jr, et al. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: the jnc 7 report. *Jama*, 289(19):2560–2571, 2003.

[28] K. J. Cios and G. William Moore. Uniqueness of medical data mining. *Artificial intelligence in medicine*, 26(1):1–24, 2002.

[29] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[30] F. S. Collins and H. Varmus. A New Initiative on Precision Medicine. *The New England Journal of Medicine*, 372(9):793–795, 2015.

[31] J. Coresh, B. C. Astor, T. Greene, G. Eknoyan, and A. S. Levey. Prevalence of chronic kidney disease and decreased kidney function in the adult us population: Third national health and nutrition examination survey. *American journal of Kidney diseases*, 41(1):1–12, 2003.

[32] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):557–562, 2011.

[33] R. A. Deyo, D. C. Cherkin, and M. A. Ciol. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *Journal of Clinical Epidemiology*, 45(6):613–619, 1992.

[34] S. Dua, U. R. Acharya, and P. Dua. *Machine Learning in Healthcare Informatics*, volume 56 of *Intelligent Systems Reference Library*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[35] S. Earley. The Promise of Healthcare Analytics. *Computing Edge*, (June):27–29, 2015.

[36] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Laleci. A survey and analysis of Electronic Healthcare Record standards. *ACM Computing Surveys*, 37(4):277–315, 2005.

[37] N. Esfandiary, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar. Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, 41(9), Jul. 2014.

[38] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2004.

[39] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[40] D. Feeny, W. Furlong, M. Boyle, and G. W. Torrance. Multi-attribute health status classification systems. *Pharmacoeconomics*, 7(6):490–502, 1995.

[41] M. M. Finucane, G. A. Stevens, M. J. Cowan, G. Danaei, J. K. Lin, C. J. Paciorek, G. M. Singh, H. R. Gutierrez, Y. Lu, A. N. Bahalim, et al. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9· 1 million participants. *The Lancet*, 377(9765):557–567, 2011.

[42] A. Fisher and D. Burke. Critical Care Scoring Systems. In *Contemporary Coloproctology*, chapter 35, pages 513–528. Springer London, 2012.

[43] A. Fisher and D. Burke. Critical care scoring systems. In *Contemporary Coloproctology*, pages 513–528. Springer, 2012.

[44] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, 25(5):845–869, 2014.

[45] M. J. García-Zattera, T. Mutsvari, a. Jara, D. Declerck, and E. Lesaffre. Correcting for misclassification for a monotone disease process with an application in dental research. *Statistics in Medicine*, 29(30):3103–3117, Dec. 2010.

[46] V. Garla, C. Taylor, and C. Brandt. Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management. *Journal of Biomedical Informatics*, 46(5):869–875, 2013.

[47] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic. Extraction of interpretable multivariate patterns for early diagnostics. *IEEE International Conference on Data Mining*, pages 201–210, 2013.

[48] R. W. Gifford, M. Alderman, A. Chobanian, S. Cunningham, H. Dustan, C. Francis, E. Frohlich, N. Kaplan, R. Mann, M. Moser, et al. The 5th report of the joint national committee on detection, evaluation, and treatment of high blood-pressure (jnc v). *Archives of Internal Medicine*, 153(2):154–183, 1993.

[49] P. A. Glare and C. T. Sinclair. Palliative medicine review: prognostication. *Journal of palliative medicine*, 11(1):84–103, 2008.

[50] P. Greenberg, C. Cox, M. M. LeBeau, P. Fenaux, P. Morel, G. Sanz, M. Sanz, T. Vallespi, T. Hamblin, D. Oscier, K. Ohyashiki, K. Toyama, C. Aul, G. Mufti, and J. Bennett. International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood*, 89(6):2079–88, 1997.

[51] D. a. Grimes and K. F. Schulz. Uses and abuses of screening tests. *Lancet*, 359(9309):881–884, 2002.

[52] A. Guillory and J. Bilmes. Label Selection on Graphs. In *Neural Information Processing Systems*, pages 1–9, 2009.

[53] E. J. Gumbel. *Statistics of Extremes*. Courier Dover Publications, 2012.

[54] R. B. Haynes, K. A. McKibbon, N. L. Wilczynski, S. D. Walter, and S. R. Werre. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ (Clinical research ed.)*, 330(7501):1179, may 2005.

[55] Health Promotion Administration, Ministry of Health and Welfare, Taiwan. 2014 Annual Report of Health Promotion Administration Health, 2014. Available from `http://www.hpa.gov.tw/English/ClassShow.aspx?No=201412160001`.

[56] M. Herland, T. M. Khoshgoftaar, and R. Wald. A review of data mining using big data in health informatics. *Journal Of Big Data*, 1(1):2, 2014.

[57] W. R. Hersh, M. G. Weiner, P. J. Embi, J. R. Logan, P. R. Payne, E. V. Bernstam, H. P. Lehmann, G. Hripcsak, T. H. Hartzog, J. J. Cimino, and J. H. Saltz. Caveats for the use of operational

electronic health record data in comparative effectiveness research. *Medical Care*, 51(8 Suppl 3):S30–7, 2013.

[58] T. L. Higgins, D. Teres, W. Copes, B. Nathanson, M. Stark, and A. Kramer. Updated mortality probability model (mpm iii). *Chest*, 128(4):348S–348S, 2005.

[59] D. Hristovski, T. Rindflesch, and B. Peterlin. Using literature-based discovery to identify novel therapeutic approaches. *Cardiovascular & hematological agents in medicinal chemistry*, 11(1):14–24, 2013.

[60] H. Huang, J. Li, and J. Liu. Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis. *Expert Systems with Applications*, 39(3):2314–2320, 2012.

[61] T. Hwang and R. Kuang. A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery. *SIAM International Conference on Data Mining*, pages 583–594, 2010.

[62] J. Iavindrasana, G. Cohen, a. Depeursinge, H. Müller, R. Meyer, and a. Geissbuhler. Clinical data mining: a review. *Yearbook of medical informatics*, pages 121–33, jan 2009.

[63] Z. Jakab, E. Comparative, and T. Eu. Periodic health examination: A brief history and critical assessment. *Eurohealth*, 15(4):16–20, 2009.

[64] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen. Application of classification techniques on development an early-warning system for chronic illnesses. *Expert Systems with Applications*, 39(10):8852–8858, Aug. 2012.

[65] M. Ji. GNetMine. `https://github.com/rackingroll/HetePathMine/blob/master/GNetMine.m`, 2010.

[66] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1298–1306, 2011.

[67] M. Ji, Q. He, J. Han, and S. Spangler. Mining strong relevance between heterogeneous entities from unstructured biomedical data. *Data Mining and Knowledge Discovery*, 29(4):976–998, 2015.

[68] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao. Graph regularized transductive classification on heterogeneous information networks. *Machine Learning and Knowledge Discovery in Databases*, 6321 LNAI:570–586, 2010.

[69] R. Kaplan, J. Bush, and C. Berry. Health status: types of validity and the index of well-being. *Health Services Research*, 11(4):478–507, 1976.

[70] A. Kastrin, T. C. Rindflesch, and D. Hristovski. Link Prediction on the Semantic MEDLINE Network. *Discovery Science*, pages 135–143, 2014.

[71] M. T. Keegan, O. Gajic, and B. Afessa. Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. *Chest*, 142(4):851–858, 2012.

[72] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, 2012.

[73] J. Kim and H. Shin. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *Journal of the American Medical Informatics Association : JAMIA*, 20(4):613–618, 2013.

[74] H. C. Koh and G. Tan. Data mining applications in healthcare. *Journal of Healthcare Information ManagementVol*, 19(2):65, 2011.

[75] E. Kontio, A. Airola, T. Pahikkala, H. Lundgren-Laine, K. Junttila, H. Korvenranta, T. Salakoski, and S. Salanterä. Predicting patient acuity from electronic patient records. *Journal of Biomedical Informatics*, 51:8–13, 2014.

[76] M. A. Krall, A. V. Gundlapalli, and M. H. Samore. Big Data and Population-Based Decision Support. In *Clinical Decision Support*, pages 363–382. Elsevier, second edi edition, 2014.

[77] T. Lancet. Chronic kidney disease: refining diagnosis and management. *The Lancet*, 384(9941):378, 2014.

[78] J. E. Leonard, J. B. Colombe, and J. L. Levy. Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics*, 18(11):1515–1522, 2002.

[79] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[80] Y. Li and J. C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010.

[81] C.-J. Lin, Hsuan-Tien and R. C. Weng. A note on Platts probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, Aug. 2007.

[82] L. Lin, P. J.-H. Hu, and O. R. Liu Sheng. A decision support system for lower back pain diagnosis: Uncertainty management and clinical evaluations. *Decision Support Systems*, 42(2):1152–1169, Nov. 2006.

[83] B. Liu, W. S. Lee, P. S. Yu, and X. Li. Partially Supervised Classification of Text Documents. In *International Conference on Machine Learning*, pages 387–394, 2002.

[84] C. Liu, F. Wang, J. Hu, and H. Xiong. Temporal Phenotyping from Longitudinal Electronic Health Records. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 705–714, 2015.

[85] N. Liu, Z. Lin, J. Cao, Z. Koh, T. Zhang, G. B. Huang, W. Ser, and M. E. H. Ong. An intelligent scoring system and its application to cardiac arrest prediction. *IEEE Transactions on Information Technology in Biomedicine*, 16(6):1324–1331, 2012.

[86] K. l. LT, J. r. KJ, G. n. j. L. C, and G. t. PC. General health checks in adults for reducing morbidity and mortality from disease (Review). *Cochrane Database of Systematic Reviews*, (10), 2012.

[87] A. Malossini, E. Blanzieri, and R. T. Ng. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics (Oxford, England)*, 22(17):2114–2121, Sept. 2006.

[88] Y. Mao, W. Chen, Y. Chen, C. Lu, S. Louis, M. Kollef, and T. C. Bailey. An integrated data mining approach to real-time clinical monitoring and deterioration warning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1140–1148, Beijing, China, 2012. ACM.

[89] L. Mirel and K. Carper. Statistical Brief - Trends in Health Care Expenditures for the Elderly, Age 65 and Over. Technical report, Agency for Healtcare Research and Quality, 2014.

[90] M. S. Mohktar, S. J. Redmond, N. C. Antoniades, P. D. Rochford, J. J. Pretto, J. Basilakis, N. H. Lovell, and C. F. McDonald. Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data. *Artificial Intelligence in Medicine*, 63(1):51–59, 2015.

[91] R. P. Moreno, P. G. H. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, and J. R. Le Gall. SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine*, 31(10):1345–1355, 2005.

[92] H. Neuvirth, M. Ozery-Flato, J. Hu, J. Laserson, M. S. Kohn, S. Ebadollahi, and M. Rosen-Zvi. Toward personalized care management of patients at risk: the diabetes case study. In *SIGKDD*, pages 395–403, California, USA, 2011. ACM.

[93] A. B. Newman, R. M. Boudreau, B. L. Naydeck, L. F. Fried, and T. B. Harris. A physiologic index of comorbidity: relationship to mortality and disability. *The journals of gerontology. Series A, Biological sciences and medical sciences*, 63(6):603–9, jun 2008.

[94] A. Y. Ng. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78, Alberta, Canada, 2004. ACM.

[95] L. T. Nguyen, M. Zeng, P. Tague, and J. Zhang. I Did Not Smoke 100 Cigarettes Today! Avoiding False Positives in Real-World Activity Recognition. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 1053–1063, 2015.

[96] Q. Nguyen, H. Valizadegan, and M. Hauskrecht. Learning classification models with soft-label information. *Journal of the American Medical Informatics Association : JAMIA*, 21(3):501–8, 2014.

[97] T. P. Nguyen and T. B. Ho. Detecting disease genes based on semi-supervised learning and protein-protein interaction networks. *Artificial Intelligence in Medicine*, 54(1):63–71, 2012.

[98] F. Nie. GGSSL code. `https://sites.google.com/site/feipingnie/publications`, 2010.

[99] F. Nie, S. Xiang, Y. Liu, and C. Zhang. A general graph-based semi-supervised learning with novel class discovery. *Neural Computing and Applications*, 19:549–555, 2010.

[100] J. PA, O. S, C. BL, and et al. 2014 evidence-based guideline for the management of high blood pressure in adults: Report from the panel members appointed to the eighth joint national committee (jnc 8). *JAMA*, 311(5):507–520, 2014.

[101] S. Palit, M. Chonchol, A. K. Cheung, J. Kaufman, G. Smits, and J. Kendrick. Association of bp with death, cardiovascular events, and progression to chronic dialysis in patients with advanced kidney disease. *Clinical Journal of the American Society of Nephrology*, 10(6):934–940, 2015.

[102] S. Pan, J. Wu, and X. Zhu. Cogboost: Boosting for fast cost-sensitive graph classification. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):2933–2946, Nov 2015.

[103] A. Pantelopoulos and N. G. Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 40(1):1–12, 2010.

[104] M. Pechenizkiy, a. Tsymbal, S. Puuronen, and O. Pechenizkiy. Class Noise and Supervised Learning in Medical Domains: The Effect of Feature Extraction. *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pages 708–713, 2006.

[105] M. Porta. *A Dictionary of Epidemiology*. Oxford University Press, New York, 5th edition, 2008.

[106] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang. A relative similarity based method for interactive patient risk prediction. *Data Mining and Knowledge Discovery*, 29(4):1070–1093, 2015.

[107] M. Rantalainen and C. Holmes. Accounting for Control Mislabeling in CaseControl Biomarker Studies. *Journal of Proteome Research*, pages 5562–5567, 2011.

[108] M. Ressing, M. Blettner, and S. J. Klug. Data analysis of epidemiological studies: part 11 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt international*, 107(11):187–92, 2010.

[109] F. Riaz, A. Hassan, S. Rehman, I. K. Niazi, and K. Dremstrup. Emd-based temporal and spectral features for the classification of eeg signals using supervised learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(1):28–35, 2016.

[110] S. Rosset, C. Perlich, G. Świrszcz, P. Melville, and Y. Liu. Medical data mining: insights from winning two competitions. *Data Mining and Knowledge Discovery*, 20(3):439–468, 2010.

[111] M. J. Rothman, S. I. Rothman, and D. B. Rothman. Systems and methods for providing a health score for a patient, Mar. 2013. US Patent 8,403,847.

[112] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[113] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li. Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus. *IEEE Transactions Knowledge and Data Engineering*, 27(1):130–141, 2015.

[114] P. Srinivasan and B. Libbus. Mining medline for implicit links between dietary substances and diseases. *Bioinformatics*, 20(suppl 1):i290–i296, 2004.

[115] E. K. Sumaili, E. P. Cohen, C. V. Zinga, J.-M. Krzesinski, N. M. Pakasa, and N. M. Nseka. High prevalence of undiagnosed chronic kidney disease among at-risk population in kinshasa, the democratic republic of congo. *BMC nephrology*, 10(1):1, 2009.

[116] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2009.

[117] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*. Pearson Addison Wesley, Boston, 2005.

[118] R. M. Tanner, D. A. Calhoun, E. K. Bell, C. B. Bowling, O. M. Gutiérrez, M. R. Irvin, D. T. Lackland, S. Oparil, W. McClellan, D. G. Warnock, et al. Incident esrd and treatment-resistant hypertension: the reasons for geographic and racial differences in stroke (regards) study. *American Journal of Kidney Diseases*, 63(5):781–788, 2014.

[119] F. Tedla, A. Brar, R. Browne, and C. Brown. Hypertension in chronic kidney disease: navigating the evidence. *International journal of hypertension*, 2011, 2011.

[120] T. Tran, D. Phung, W. Luo, R. Harvey, M. Berk, and S. Venkatesh. An integrated framework for suicide risk prediction. In *SIGKDD*, pages 1410–1418, Chicago, USA, 2013. ACM.

[121] T. Tran, D. Phung, W. Luo, and S. Venkatesh. Stabilized sparse ordinal regression for medical risk stratification. *Knowledge and Information Systems*, 43(3):555–582, mar 2015.

[122] F. Wang, P. Zhang, B. Qian, X. Wang, and I. Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 145–154, New York, USA, 2014. ACM.

[123] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, 2013.

[124] X. Wang, F. Wang, J. Wang, B. Qian, and J. Hu. Exploring patient risk groups with incomplete knowledge. *IEEE International Conference on Data Mining*, pages 1223–1228, 2013.

[125] J. M. Wei, S. Q. Wang, and X. J. Yuan. Ensemble rough hypercuboid approach for classifying cancers. *IEEE Transactions on Knowledge and Data Engineering*, 22(3):381–391, 2010.

[126] C. P. Wen, T. Y. D. Cheng, M. K. Tsai, Y. C. Chang, H. T. Chan, S. P. Tsai, P. H. Chiang, C. C. Hsu, P. K. Sung, Y. H. Hsu, and S. F. Wen. All-cause mortality attributable to chronic kidney disease: a prospective cohort study based on 462 293 adults in Taiwan. *Lancet*, 371(9631):2173–2182, June 2008.

[127] J. Wiens, E. Horvitz, and J. V. Guttag. Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task. In *Neural Information Processing Systems*, pages 476–484, 2012.

[128] M. Woodward. *Epidemiology: study design and data analysis*. CRC Press, 2013.

[129] Z. Xing, J. Pei, and E. Keogh. A brief survey on sequence classification. *SIGKDD Explor. Newsl.*, 12(1):40–48, Nov. 2010.

[130] P. Yang, X. Li, H. N. Chua, C. K. Kwoh, and S. K. Ng. Ensemble positive unlabeled learning for disease gene identification. *PLoS ONE*, 9(5), 2014.

[131] P. Yang, X. L. Li, J. P. Mei, C. K. Kwoh, and S. K. Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 2012.

[132] Y. Yang, Z. Ma, A. Hauptmann, and N. Sebe. Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia*, 15(3):661–669, April 2013.

[133] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. G. Hauptmann. How related exemplars help complex event detection in web videos? In *IEEE International Conference on Computer Vision (ICCV'13)*, pages 2104–2111. IEEE, 2013.

[134] J.-Y. Yeh, T.-H. Wu, and C.-W. Tsao. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems*, 50(2):439–448, Jan. 2011.

[135] S.-I. Yi, B.-R. So, C.-S. Lee, S.-J. Lee, S.-K. Park, B.-K. Park, and I.-W. Chung. Classification of Health Grade Using Bio-Check Unit and Health Index. *Journal of Biomechanical Science and Engineering*, 6(3):148–159, 2011.

[136] I. Yoo and X. Hu. A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries - JCDL '06*, page 220, 2006.

[137] N. Zemmal, N. Azizi, N. Dey, and M. Sellami. Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification. *Journal of Medical Imaging and Health Informatics*, 6(1):53–62, 2016.

[138] F. Zhang, Y. Song, and W. Cai. A ranking-based lung nodule image classification method using unlabeled image knowledge. *IEEE 11th International Symposium on Biomedical Imaging*, pages 1356–1359, 2014.

[139] L. Zhang, F. Wang, L. Wang, W. Wang, B. Liu, J. Liu, M. Chen, Q. He, Y. Liao, X. Yu, et al. Prevalence of chronic kidney disease in china: a cross-sectional survey. *The Lancet*, 379(9818):815–822, 2012.

[140] M. Zhao, R. H. M. Chan, T. W. S. Chow, and P. Tang. Compact Graph based Semi-Supervised Learning for Medical Diagnosis in Alzheimer's Disease. *IEEE Signal Processing Letters*, 21(10):1192–1196, 2014.

[141] Y. Zhao, G. Wang, X. Zhang, J. X. Yu, and Z. Wang. Learning phenotype structure using sequence model. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):667–681, 2014.

[142] J. Zhi-chun. Correctly recognizing and treating the defects of evidence-based medicine practice. *Journal of Evidence-based Medicine*, 13(5):310, 2013.

[143] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Sch. Learning with Local and Global Consistency. *Advances in Neural Information Processing Systems*, 1:595–602, 2003.

[144] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *International Conference on Machine Learning*, 20(2):912–919, 2003.

[145] J. E. Zimmerman, A. a. Kramer, D. S. McNair, and F. M. Malila. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Critical care medicine*, 34(5):1297–1310, 2006.

Appendix

Awards During Candidature

- Xue Li, Weitong Chen, Sayan Unankard, Ling Chen, Xiaofang Zhou, Shazia Sadiq. **The Best Use of Open Data Award**. WikiQueensland- a big data fusion system, in *The 2014 Premier's Awards for Open Data*, Queensland, Australia, 2014.
- Xue Li, Weitong Chen, Sayan Unankard, Ling Chen, Xiaofang Zhou, Shazia Sadiq. **Microsoft StartUp Q Award**. WikiQueensland- a big data fusion system, in *The 2014 Premier's Awards for Open Data*, Queensland, Australia, 2014.
- Ling Chen, Xue Li, Quan Z. Sheng, and Mohamed A. Sharaf. **Best Research Poster Awards - Joint First Prize**, Mining Personal Health Index from Annual Geriatric Medical Examinations, in *The 2014 Australasia Database Conference PhD School in Big Data*, 2014.
- Ling Chen, **Richard Jago Memorial Prize** 2014, The University of Queensland.
- Ling Chen, UQ Graduate School sponsorship for attending the 2014 C9-Go8 Forum on Global challenges of Ageing Populations, Sydney.