KNOWLEDGE DISCOVERY FOR HEALTH RISK PREDICTION

A Thesis submitted by

Thuan Pham

BEng, MSc

For the award of

Doctor of Philosophy

2020

Abstract

Improving the accuracy of the diagnosis of disease can help to increase the quality of healthcare. Many researchers have developed classification models to support healthcare practitioners to make accurate diagnoses, avoiding the need to rely on their experience base diagnose diseases. However, these models are currently based on datasets collected from healthcare data including medical history. As a result, the reliability and accuracy of predicting results of the diagnosis, are limited.

Following the goal of improving the accuracy of health risk prediction, this thesis concentrates on the classification of tasks through mining healthcare data. The study suggests several frameworks and algorithms to develop classification models. In addition, challenges of extracting useful information and processing data noise from the real dataset are addressed as a way of learning models. Classification models are developed based on well-proven medical data sources. By using medical evidence, the study aims to improve the accuracy of classification for health risk prediction.

The first contribution of this thesis is an innovation of building a binary classification model to predict patients' risks. The second contribution of this dissertation is to build a medical knowledge base to support classification models for improving the reliability and accuracy of the model. The third significant contribution of the thesis provides a framework for building a predictive model within multiple diseases.

Certification of Thesis

This thesis is the work of Thuan Pham except where otherwise acknowledged. The work is original and has not previously been submitted for any other award, except where acknowledged.

Principal Supervisor: Associate Professor Xiaohui Tao

Associate Supervisor: Associate Professor Ji Zhang

Associate Supervisor: Professor Jianming Yong

Student and supervisors' signatures of endorsement are held at USQ.

Acknowledgements

My thanks to these individuals, who have dedicated assistance, motivation, and encouragement to me during my PhD research. I would not be able to achieve the current study without their support. In particular, I would like to express my thanks and sincere appreciation to: Associate Professor Xiaohui Tao, my principal supervisor, for his patience and motivation to support my PhD study. With immense knowledge, his critical analysis, quick response and guidance assisted me to build up my research foundation and to complete my research. My thesis would not have been achievable without his assistance. Associate Professor Ji Zhang, my associate supervisor, for his scientific advice and useful remarks for my PhD study. Professor Jianming Yong, my associate supervisor, to whom I am grateful for plentiful comments of my research and opportunities in teaching as well. Dr Barbara Harmes, who has assisted me in proofreading all the journal articles as well as the dissertation. My parents and friends, who have always encouraged me throughout the journey of my PhD.

Contents

List of Figures

List of Tables

List of Abbreviations

MeSH	**Me**dical **S**ubject **H**eadings
ICD	**I**nternational **C**lassification of **D**iseases
NLM	**N**ational **L**ibrary of **M**edicine
UMLS	**U**nified **M**edical **L**anguage **S**ystem
NHANES	**N**ational **H**ealth **A**nd **N**utrition **E**xamination **S**urvey
NAMCS	**N**ational **A**mbulatory **M**edical **C**are **S**urveys
GHE	**G**eneral **H**ealth **E**xamination
EHR	**E**lectronic **H**ealth **R**ecord
HIG	**H**eterogeneous **I**nformation **G**raph
KB-HIG	**K**nowledge **B**ase **H**eterogeneous **I**nformation **G**raph
SHG-Health	**S**emi-supervised **H**eterogeneous **G**raph on **Health**
KB-SHG-Health	**K**nowledge **B**ase **S**emi-supervised **H**eterogeneous **G**raph-based on **Health**
MulGRaL	**Mul**ti-label **G**raph **Ra**nking **L**earning
Mul-SHG-Health	**Mul**ti-label **S**emi-supervised **H**eterogeneous **G**raph on **Health**

Chapter 1

Introduction

1.1 Background

With the rapid development of technology in medicine, the quality of healthcare is receiving closer attention and many countries have invested significant amounts of money in public healthcare. One of the most significant is the United States government that in 2011 disbursed $414.3 billion on healthcare (Mirel and Carper [2014]). At the same time, the United States government endeavoured to enhance precision medicine by utilising increasingly large amounts of available health data (Collins and Varmus [2015]). They aimed to create a valid database for analysing and justifying the health risk status of patients. Some years later, the Australian government expended $170 billion from 2015 to 2016 for healthcare (Inacio et al. [2019]), highlighting the importance of spending money and effort to improve the quality of healthcare. However, improving healthcare is challenging; finance is necessary to support and improve healthcare, but technologies are needed to assist in achieving the goal of improvement in healthcare quality.

In healthcare, data mining is used to check treatment effectiveness, health

customer relationship management, patient care, and to counter fraud and abuse (Koh et al. [2011]). Much work has been done in disease risk assessment, with a focus on supporting medical practitioners to make safe and effective clinical decision. Massive medical datasets contain wealth domain knowledge that can help physicians in decision-making. There is strong evidence that clinical decisions based on risk assessment may improve disease management (James et al. [2014], Hunink et al. [2014]). Data mining techniques have increased as a way of predicting diseases (Chang et al. [2011], Huang et al. [2012]). Some researchers also have developed predictive models for the classification of clinical risks (Kim et al. [2014], Sabibullah et al. [2013]). These studies have helped to improve accuracy in disease assessment and reduce errors in disease treatment.

In developing systems for health risk prediction, exploring all relationships between medical history and medical condition plays an important role to improve the performance of classification models. Therefore, many researchers (Zhang et al. [2018a], Shah et al. [2019], Han et al. [2019]) have focused on mining label-specific features and label correlations to build classification models as well as boost the performance of classification models. To explore label correlation effectively, a heterogeneous information graph is used to mine these relationships between label correlations. A large number of researchers (Chen et al. [2016a], Xiong et al. [2018], Lei and Zhang [2019], Wang et al. [2020]) has considered using a graph to develop classification models. These studies have brought significant improvements in enhancing the accuracy of health risk prediction.

While data mining may be beneficial for supporting healthcare, some

challenges need to be addressed at the same time in the process of developing models to support the assessment of health risks. Support models are often built based on observational data collected from patient conditions described in electronic health records. However, these datasets still have challenges and require many preprocessing steps before becoming useful data to apply in building models. The first challenge is the heterogeneity of the data. Datasets usually contain a large number of different data types combined from such evidence as physical examinations, personal habits, and lab tests, etc. These types of data are related to different parts of the human body and various aspect of a patient's health status. These factors are critical and need to be handled effectively to achieve classification models with the highest performance. The dataset with an insufficient number of labelled samples also influences the training model. If most of the target labels used for assessment are negative cases, the performance of models will have less reliability. Moreover, non-completion of the data collection also brings challenges to the development of effective models. If a dataset has too much noise, this will negatively affect the accuracy and quality of the model. These challenges must be overcome to achieve classification models with effective results of prediction.

In the field of healthcare, especially for diagnosis, evidence and reliability are extremely important. Therefore, the evidence-based medicine approach is expected to be used for optimising decision-making. It aims to ensure that the clinician's opinion relies not only on available knowledge from the scientific literature but also on local knowledge mined from individual observational data. The approach helps minimise risks during the diagnosis process. Therefore, disease prediction models are more useful if

they can integrate evidence to ensure the reliability of that model. In the field of medical research, Medical Literature Analysis and Retrieval System Online (MEDLINE) is one of the most significant data sources related to scientific literature. It is a metadata repository of biomedical abstracts and uses Medical Subject Headings (MeSH) to manually index publications from the National Library of Medicine (NLM). Wang et al. [2017a] showed that the integration of medical knowledge has a strong ability to improve information retrieval performance for medical informatics applications, like recommender systems or ontology learning. By relying on the personalised medical profile of the patients and knowledge bases (MEDLINE), researchers can provide evidence-based decision-making support to healthcare practitioners, which promises to improve the quality of healthcare services. Data mining could be the best approach to help expand the applications to the healthcare industry. It is expected to have a significant effect on decision support systems and improving the quality of healthcare institutes.

1.2 Statement of the Problem

The main problem addressed in this research is the classification of healthcare data. The research tries to improve the disease diagnosis result by using data mining and machine learning techniques. The research thesis provides evidence-based decision-making support to doctors, physicians and healthcare practitioners and helps them reduce human errors. The human brain has limits; medical knowledge changes over time. Doctors and physicians may find it difficult to avoid human errors when they rely on their experience for medical advice, diagnosis or treatment. Experience-based decisions may lead to the problem that some critical cases are overlooked. In

contrast, data mining in healthcare can help cover the overlooked areas because it does not have the limitations mentioned above. Data mining allows researchers to work with data collected from a massive number of patients, a number that is more than any doctor ever treats. As a result, data mining can provide high-quality evidence covering as many possibilities as possible to support doctors' decision-making by knowledge discovery in healthcare data.

In terms of technology and knowledge, in order to provide an evidence base for medical decisions, the study uses data mining and machine learning to develop classification models that are able to support medical advice through the evidence base of medicine. Having evidence from data analysis helps doctors to be more confident in predicting a patient's health status as being healthy or unhealthy through a binary classification, as well as assessing multiple diseases through a multi-label classification. Furthermore, a classification model not only helps doctors categorize patients as being healthy or unhealthy but also gives them suggestions about what kind of diseases they are suffering from.

Although evidence-based medicine plays an important role in developing classification models for health risk prediction, integrating knowledge into classification models is not an easy task. It requires a lot of effort and time to process these challenges. Moreover, building classification models always is challenging because of the heterogeneity of the data, imbalance of the data or missing data. To deal with the task of predicting health risk status through evidence based diagnosis, research needs to present some of the effective methods to solve these issues. To achieve the research goals of this thesis, there are four research questions that will be addressed as follows:

Q1) What are the essential underlying patterns from a patient's observational data that may lead to the conclusion that the patient's health status is healthy or unhealthy?

The research based on the observational data from the National Health and Nutrition Examination Survey (NHANES)[1] dataset to develop classification models, is able to predict whether a patient is healthy or unhealthy. The NHANES dataset has a thousand attributions about the characteristics of a patient. For example, a patient is suffering from liver cancer. A question may be asked whether people who smoke every day, or drink alcohol three times per week are likely to suffer from liver cancer. Another issue is how to define the relationship between disease and attributions. It is clear that identifying these underlying patterns is a challenging task.

Q2) Given a set of diseases, what are the identifiers in a patient's observational data that can be used to separate one disease from others? Not only is this study aimed to identify a patient affected by a disease, but it is also about whether the patient suffers multiple diseases. Identifying a patient affected by many diseases is much more complicated than identifying a specific disease. Classification models need to identify factors for the determination of different diseases. How to predict different diseases is another challenge in this study. The classification model is needed to find out the similar and different relationships between diverse diseases from the numerous inputs of attributes. The impact of attributes on different diseases also needs to be assessed to achieve a useful classification model.

[1] https://www.cdc.gov/nchs/nhanes/index.htm

Q3) How can the findings discovered in Q1 be used in a classification model to diagnose a patient's health status based on evidence?

Given $\mathcal{P} = \{p_1, p_2, ..., p_n\}$ is a set of patients and $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ is a set of attributes. Each patient p links with their attributes such as age, weight, lab test results, habits. $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ is a set of diseases that is associated with patients in $\mathcal{P}$. A classification model, which is developed based on $\mathcal{P}$, $\mathcal{A}$ and $\mathcal{D}$, may have limited information to assess whether a patient is healthy or unhealthy. Discovering underlying patterns from a dataset to identify the connections between attributes and different diseases is a challenge in this study. To improve the accuracy and reliability of the classification model, the research endeavours to discover more relationships in $\mathcal{A}$ by mining knowledge from the MEDLINE corpus. How to discover connections between observational data and MEDLINE is the major challenging task in this study.

Q4) How can the findings discovered in Q2 be used in a multi-label classification model to diagnose which diseases a patient is suffering from?

Identifying the attributes that cause each disease plays an essential role in forming a predictive model for that disease. Some previous work has succeeded in this issue. However, it is more challenging in terms of determining the connection of attributions to multiple diseases. A patient in $\mathcal{P}$ may associates with one or more disease in $\mathcal{D}$. Therefore, discovering underlying connections among patterns in $\mathcal{A}$ and $\mathcal{D}$ is an important task in developing the multi-label classification model. Besides, identifying the relevance among diseases in $\mathcal{D}$ is expected to improve the accuracy

of the multi-label classification model. By identifying connections between attributes and different diseases, a multi-label classification model can be developed for support practitioners in predicting multiple diseases.

1.3 Research Aims and Objectives

Based on the research problem, the thesis aims to develop innovative classification models to assist practitioners in medical diagnosis. The classification models mine attributes of a patient from the observational data in NHANES. Besides, these models can be expected to integrate knowledge in the medical domain that has been evaluated by experts to increase the accuracy and reliability of predicting results. In the past, researchers relied only on the attribute of the patients but were not interested in using medical knowledge to evaluate and verify the useful attribute of the patients from the collected data. The proposed model is expected to improve the quality of clinical evidence-based decisions as well as reduce the time and cost incurred during the diagnosis by practitioners. In particular, this thesis achieves the following objectives:

(I) To develop a binary classification model for assessing a person's health status to see whether he (she) is healthy or not. Based on the observational data, this study analyses and processes the data in the dataset to find correlations between a patient's attributes and the diseases that a patient is likely to have. A model then is developed to predict diseases for patients by using these findings. This objective is designed to deal with the research question Q1 in Section 1.2.

(II) To integrate the knowledge into the classification model for generating an evidence-based model to support practitioners. This objective requires an innovative framework to exploit how to integrate the knowledge base into classification models. This research uses MEDLINE and MeSH as an evidence-based source to adopt to the classification model. A lot of effort may be expected to transform the knowledge in MEDLINE and MeSH for acquiring a deep understanding of using the observational data of NHANES. This task is challenging in presenting both data and knowledge in the data analysis format. Mapping the observational data and the knowledge base is expected to improve the performance of classification. A comparison between models within and without knowledge is required to evaluate how the effect of the knowledge base compares to the traditional classification model. This objective is designed to address the research question Q3 specific in Section 1.2.

(III) To design a multi-label classification model for predicting possible diseases for a patient. The multi-label classification model is more complicated than a binary classification model. The study is not only to mine underlying connections between symptoms and disease but also to mine among diseases. Moreover, integrating the knowledge into the multi-label classification model needs more time and cost. Assuming connections among patterns can be discovered, however, applying evidences of medicine to improve these connections which help to advance the performance of classification model still are remains challenging. Therefore, an innovative approach is needed to solve the

classification of multiple diseases. This objective is designed to conduct the research question Q2 and Q4 in Section 1.2.

1.4 Scope and Limitations

The scope of the thesis is on mining the observational data from NHANES, which was collected from 2013 to 2014 to conduct the experimental results of the proposed thesis. This research is limited by the number of diseases due to the limit of available data. The amount of missing data for some datasets are often huge, so the results of analysis and process data to evaluate for some diseases may decrease accuracy. Therefore, this study develops a model to predict diseases that have less than fifty per cent of missing data.

In addition, the limitation of this study is the number of patients of the healthcare dataset is a constant, which leads to data flexibility limitation. A dynamic study is more complex and is expected to explore in future work. The time-series data of healthcare data also has not been considered in this study. The time series plays an important role in determining the outcome of the prediction. This may have an impact both directly and indirectly on conditions from time to time.

1.5 Organisation of the Thesis

This section presents the organisation of the thesis. At the highest level, this research builds effective classification models for healthcare decision support to help physicians to avoid human errors. The flow structure of the thesis is shown in Figure 1.1.

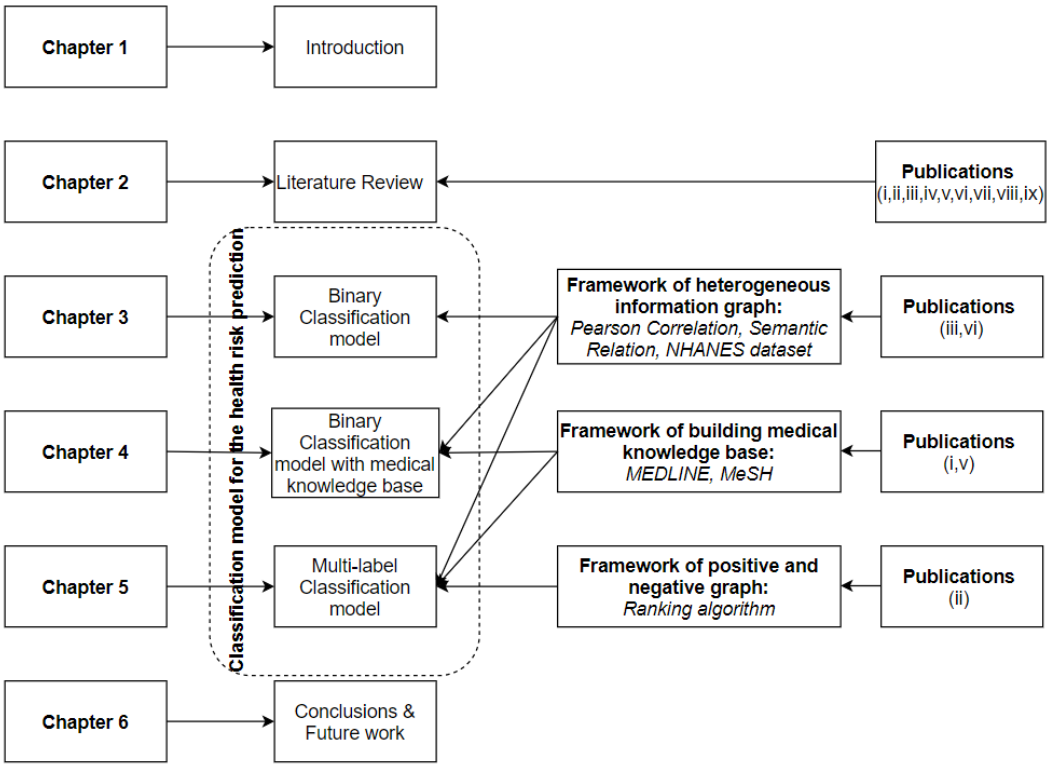


FIGURE 1.1: The flow structure of the thesis

The research background, the statement of the research problem, aim of the research and scope of the research are presented in Chapter 1. The study introduces the related works of the thesis in chapter 2. The main task of the doctoral thesis consists of three major studies, which cover the three objectives corresponding with Chapter 3, 4, and 5. Chapter 6 presents conclusions with scientific contributions and the future work.

- In Chapter 2, the research first discusses previous work about mining data, specifically in the medical domain. How do they build a classification model to support in diagnosis? What are the best methods being considered in predicting the health risk status? Then, the thesis presents research around learning classification through a knowledge

graph. Finally, the study shows related work about multi-label classification problems.

- In Chapter 3, the research focuses on the first objective. The objective is to understand the compound influences of different patterns from the observational data. The NHANES dataset is used to evaluate the proposed model. The dataset has more than 2585 attributes, which cover a wide range of information about an individual, such as personal demographics, observations, laboratory tests, or diagnostic reports. Based on the analysis of the NHANES dataset, a binary classification model is developed to predict the person's health status through the heterogeneous information graph, which can show whether a patient is healthy or not.

- Chapter 4 deals with building knowledge graph to apply to the classification model for improving the accuracy of the health risk assessment, which responds to the second objective. The study uses the evidence of medicine to recognise symptoms that have less effect on assessing the health risk status. In this project, the research uses MEDLINE as an essential source of the knowledge base to develop and evaluate the performance of the classification model. Besides, the medical subject headings, which are used for information retrieval from MEDLINE, are expected to be mined for accessing MEDLINE effectively and efficiently. The research analyses the combination of the knowledge source (MEDLINE), ontology (MeSH) and the observational data (NHANES) to develop a classification model within evidence-based medicine. To achieve a highly effective discovery of the mapping between the knowledge base and the data, a heterogeneous graph

(Ji et al. [2011a], Sun et al. [2009]) may be presented with different types of nodes to connect to multi-typed relationships of data items. A classification model, which is built based on the medical knowledge graph, is expected to achieve the high performance of classification compared to classification models without the medical knowledge graph.

- In Chapter 5, the third objective introduces a multi-label classification model to predict multiple diseases that a patient could suffer. The study processes and analyses the effect of symptoms on different diseases. The varying impacts of symptoms of diseases are then are ascertained. Later, the risk level of a symptom to different diseases can be computed. A symptom with the high-level of the influence associated with greater numbers of disease in the process of building the multi-label classification model. The symptoms belongs to different categories, such as personal demographics, observations, laboratory tests, or diagnostic reports. To improve the link between symptoms and diseases, an innovative approach called a positive and negative graph is proposed to find out connections between underlying patterns from the observational data in NHANES. A ranking algorithm then determines the risk level of a patient for learning multi-label classification.

1.6 Publications

Journal Articles

(i) **Pham, T.**, Tao, X., Zhang, J. Yong, J. Constructing a knowledge-based heterogeneous information graph for medical health status classification. Health Information Science and Systems, 8, 10 (2020).
DOI: https://doi.org/10.1007/s13755-020-0100-6

(ii) **Pham, T.**, Tao, X., Zhanag, J., Yong, J. (2020, February). Multi-Label Graph Based Disease Prediction Methods from Observational Dataset. Future Generation Computer Systems. (Under Review)

(iii) Tao, X., **Pham, T.**, Zhanag, J., Yong, J., Goh, W.P., Zhang, W., Cai, Y. (2020). Mining Health Knowledge Graph for Health Risk Prediction. World Wide Web,. (Accepted)

(iv) Zhang, J., Tan, L., Tao, X., **Pham, T.**, & Chen, B. (2020). Relational intelligence recognition in online social networks—A survey. Computer Science Review, vol.35, 100221.
DOI: https://doi.org/10.1016/j.cosrev.2019.100221

Conference Papers

(v) **Pham, T.**, Tao, X., Zhang, J., Yong, J., Zhou, X., & Gururajan, R. (2019, December). MeKG: Building a Medical Knowledge Graph by Data Mining from MEDLINE. In International Conference on Brain Informatics (pp. 159-168). Springer, Cham.
DOI: https://doi.org/10.1007/978-3-030-37078-7_16

(vi) **Pham, T.**, Tao, X., Zhang, J., Yong, J., Zhang, W., & Cai, Y. (2018, November). Mining Heterogeneous Information Graph for Health Status Classification. In 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC) (pp. 73-78). IEEE.
DOI: https://doi.org/10.1109/besc.2018.8697292

(vii) Tan, L., **Pham, T.**, Hang, K.H., & Tan, S.K. Discovering Relational Intelligence In Online Social Networks. The 31st International Conference on Database and Expert Systems Applications, Bratislava, Slovakia, September 14-17, 2020. (Submitted)

(viii) Zhang, J., Tan, L., Tao, X., **Pham, T.**, Zhu, X., Li, H., & Chang, L. (2018, November). Detecting Relational States in Online Social Networks. In 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC) (pp. 38-43). IEEE.
DOI: https://doi.org/10.1109/besc.2018.8697237

(ix) Zhang, J., Tao, T., Hang, K.H., Dokos, S., **Pham, T.**, Bai, S., & Tan, L. Predicting Events in Online Social Networks. International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, July 25-30, 2020. (Submitted)

Chapter 2

Literature Review

This chapter introduces a brief literature review concerning use of the clinical dataset as well as methods regarding the solution of classification problems. The main target of this chapter concentrates on reviewing the advantages and disadvantages of developing models for predicting health risk status. The chapter first looks at the existence of the binary classification problems: how have previous researchers analysed and processed ways to build the binary classification; how can others determine and identify the health risk status; what are the existing models for health risk prediction? Another issue related to this work focuses on construction and understanding of knowledge graphs. Finally, this chapter focuses on reviewing existing works about the problems of learning multi-label classifications.

2.1 Classification for Health Risk Analysis

2.1.1 Data Mining and Analysis in the Medical Domain

Data mining is used not only to discover underlying relationships among large observational datasets but also to summarize and provide intelligible and useful data to users (David [2007]). It is used to find unknown patterns

and trends in information (Koh et al. [2011]). Since the use of computers has become extensive in the healthcare industry, data mining has also become an essential modality in the field of health sciences. The goal of research on health information is to combine computer science and information science to improve the quality of care (Herland et al. [2014]). According to Rosset et al. [2010] and Holzinger [2016], researchers use data mining as an important tool for analyzing big data to improve healthcare services. Moreover, by using data mining techniques, healthcare professionals can predict health insurance fraud, healthcare costs, disease prognosis, disease diagnosis and disease epidemiology and accurately estimate the length of stay (LOS) in a hospital (Yoo et al. [2012]).

Ideally, data mining techniques should support intelligent data preprocessing that automatically selects the required data and eliminates the undesired data. It should also use domain knowledge of other data processes, and should also automate the knowledge discovery process. These technics lead to a better understanding and utilization of existing knowledge in data. If these problems were adequately resolved, data mining could likely become a core technology for the practice of evidence-based medicine (Yoo et al. [2012]).

2.1.2 Health Status Measurement

Health status measurement prognostication is becoming one of the most difficult challenges that health practitioners are facing. Therefore, scoring

systems play an essential role in minimizing errors caused by fatigue. In addition, these systems are widely used to support health practitioners in improving health knowledge and clinical decisions. In the area of myelodysplasia syndromes, different scoring systems are introduced, referring to different goals. For example, Miyazaki [2013] introduced an international prognostic scoring system (IPSS), which mainly focused on improving analysis of the specific impact of marrow blast percentage and depth of cytopenias. Prakash et al. [2006] suggested a Simplified Acute Physiology Score (SAPS) II scoring system, which helps physicians to make better clinical decisions by quantifying the severity of illness in the Intensive care unit area. The system introduces a method of converting the score to the possibility of a patient mortality in the hospital. In line with SAPS II, the Acute Physiology and Chronic Health Evaluation (APACHE) II scoring system (Wagner and Draper [1984]) is introduced, focusing on the systematic application of clinical judgments about the relative importance of derangement.

However, the researchers have tried not only to introduce new scoring systems but also to compare the advantages and efficiency levels among the existing scoring systems, resulting in giving a better choice of scoring systems to physicians. For example, Keegan et al. [2012] have discussed the performance of four scoring systems, including APACHE III, APACHE IV, SAPS III and Mortality Probability Model (MPM) III. Research showed that APACHE III and APACHE IV had no significant difference in distinguishing capability, and they both performed better compared with SAPS III and (MPM) III. Moreover, research also showed that complex models worked better than simple models, and the efficiency level of these models depended on how many variables being used for developing a classification

model.

2.1.3 Health Risk Prediction by Classification

Different studies have been conducted to use classifications to support health practitioners in the prediction of health risks. Yeh et al. [2011] endeavoured to apply the classification techniques to build an optimum cerebrovascular disease predictive model. In their research, three attribute input modes, T1, T2, and T3 were built, which mainly focused on building the classification models and comparing their advantages and efficiencies. On the other hand, Neuvirth et al. [2011] conducted research which mainly focused on applying state-of-the-art methods to predict the future health of patients and identify patients at high risk. In order to conduct this study, two binary classification algorithms logistic regression (LR), and k-nearest neighbour (KNN) were used.

Following this research, a new approach to machine learning was used by Nguyen et al. [2014]. Their approach used the training phase accompanying soft labels to refine the binary class information to achieve a more efficient binary classification model. In order to solve the problem of label uncertainty (label noise) in binary classification, Yang and Loog [2016] introduced a new method which was focused on using uncertainty information to improve the performance of retraining-based models. The results show that the new method provides a more efficient demonstration and can be used to reduce human labelling errors in different applications.

2.1.4 Mining Heterogeneous Graphs for Classification

The graph-based methods have brought more advantages for discovering the intrinsic characteristics of data, where the vertices and edges of a graph are taken up as model data points and their relationships, respectively (Guillory and Bilmes [2009]). Researchers have conducted different studies which aim to minimize the errors of the graph-based method. For example, the study shows that if data is presented in a heterogeneous graph, the results of mining these graphs can be significantly improved. Therefore more meaningful results can be generated from the different types among links and objects (Sun et al. [2009]). In 2010, Ji et al. [2010] conducted a study by using a classification method for heterogeneous networks. This method is called GNetMine. GnetMine uses only one classification criteria for all of the objects in the network which, it is argued, is one of the weaknesses of this method. On the other hand, however, Luo et al. [2014] discovered that the type of differences of objects in the network might have different criteria of classification. In order to improve on this weakness, they suggested a new method to minimize this drawback by providing the concept of meta paths for mining the heterogeneous graph.

In healthcare data, some researchers have taken advantage of the heterogeneous graph to discover more knowledge and improve disease diagnosis. Hwang and Kuang [2010] introduced a heterogeneous label propagation algorithm using graph-based semi-supervised learning to the discovery of disease genes. This study is based on homo-subnetworks, where links are

set up from the same types of objects to build up a heterogeneous disease-gene graph. Recently, Chen et al. [2016a] proposed a semi-supervised heterogeneous graph on the health (SHG-Health) algorithm to predict high-risk disease classes for unlabelled data. The model has contributed to a significant improvement in classification problems in healthcare data by mining knowledge from a heterogeneous graph.

2.2 Knowledge Base Learning

2.2.1 Ontology and Related Techniques

By using ontology, data can be organized as a knowledge graph, including concepts and relationships among concepts. This can help raise efficiency in design models in data mining. Many researchers have taken advantage of this technology to raise the performance of their model. Lee et al. [2006] proposed a new approach to predict more accurately the detailed concepts based on the existing concepts, at the same time using syntactic relations. This approach is useful for automatic generation of new concepts from biomedical articles, while Gao et al. [2017] tried to improve the efficiency of distance learning methods from ontology mapping and ontology similarity measuring. They suggested novel algorithms which would be presented as corresponding ontology sample data labelled and ontology sample data without a label. The results of experiments have demonstrated that this innovative ontology is more efficient and accurate in distance learning from ontology mapping and ontology similarity measures. Another approach, presenting the concept detected from a document into a graph, Ni et al. [2016] proposed a new method to measure similarly among concepts.

The approach represents concepts as continuous vectors which are utilized to accumulate similar pairwise among pairs of concepts. This study makes improvements in measuring semantics among documents.

On the other hand, Choi et al. [2014] gave a new method for improving ranking performance, called the semantic concept-enriched dependence model. The proposal shows that terms related to concepts are also crucial for improving information retrieval. The study conducts extensive experiments, including a medical literature corpus and a clinical document corpus which is more effective compared to previous work. With the focus on term space, Wang and Akella [2015] defined the notion of a concept which uses documents and queries from term space to create concept space. This study has enhanced the estimation of relevance, including decreased dimensionality of the space and remaining dependencies between the words of a concept. Similarly, to guarantee the semantic measure similarity between the medical terms, Karpagam et al. [2016] suggested a new method which combined the sources of disease concepts and biomedical resources to identify the medical terms for extending the Disease Ontology automatically.

2.2.2 Knowledge Graph

Knowledge Graph has become an essential topic in the last decade. It plays a vital role in mining data. It can help to discover hidden patterns between entities. There is an increase in building and applying a knowledge base on data mining. Xu et al. [2014a] suggested a new knowledge powered method by incorporating knowledge graphs into the learning process to encode the relationship between entities, attributes or properties of objects. This approach has improved the quality of word representations. Bordes et al.

[2011] suggested a method to learn the distributed embedding of knowledge bases. This approach has helped to generate new reasonable relations by linking raw-text as entity vectors to knowledge extraction. Similarity, Nguyen et al. [2017] investigated a method to apply semantics from raw text and knowledge resources for achieving high-level representations of documents based on both text embedding and concept-based embedding.

To use conceptual graphs effectively, Shi et al. proposed a new model to organise and integrate the textual medical knowledge into conceptual graphs. This approach provides semantic mappings between textual medical expertise and medical knowledge, which can explore complex semantics among entities in chain inferences. It also helps to detect and obtain access to valuable information from the medical domain. Moreover, based on the documents, Voskarides et al. [2015] tried to clarify the relationships among entities of knowledge graph by sentences. These sentences that refer to an entity pair were extracted and enriched through ranking.

2.3 Learning Classifiers using Knowledge Base

2.3.1 Mapping between Different Sources of Medical Terminologies

Medical terminologies are crucial not only in healthcare but also in medical research. If practitioners and researchers can have a deep understanding of terminologies in distinguishing between terms and concepts, it can help reduce the limit of human errors. Moreover, combining these terms and concepts of different sources can lead to more semantic networks of these

data. By mapping both Medical Subject Headings (MeSH) and the International Classification of Diseases (ICD-10) bases using prescription coding, Pereira et al. [2006] obtained 68% of recall, which facilitated the task of coding patient information. These two sources were extracted from the Unified Medical Language System (UMLS) Metathesaurus.

Approximately 34% of Metathesaurus strings are identified from the titles and abstracts of the biomedical literature in MEDLINE by (Srinivasan et al. [2002]). Taking advantage of this discovery, Schriml et al. [2018] built an ontology of human disease that organised concepts and terms related to the concepts of disease systems. This ontology was generated by extending and integrating the cross-mapping resources of MeSH, ICD, Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT). The knowledge presented in an ontology can be more useful in discovering the semantic network of diseases. Cross-terminology mapping of these sources to create a terminological medication system was also conducted by Saitwal et al. [2012]. Standardized biomedical terminologies can help the growing amount of research, and increase the amount of clinical and public health data. Moreover, Zhang et al. [2016] combined the MeSH terms and the UMLS concepts to improve the retrieval performance of the relevant biomedical documents from MEDLINE. 43.3% of improvement was achieves compared to the other approach without using MeSH and UMLS.

2.3.2 Knowledge Discovery in Medical Corpus

Scientific literature has significant contributions in improving precise knowledge. Discovering precise knowledge can help to strengthen the accuracy of decision support systems for the healthcare industry. Many researchers

have tried to implement different learning approaches and a variety of system applications to improve the clinical system. A model with two levels of self-supervision has been established concerning extraction using the knowledge base MEDLINE and Unified Medical Language System (UMLS) (Banuqitah et al. [2016]). In contrast, Jiang et al. [2017] suggested a three-layer knowledge base model that raised the precision in system prediction and provided more opportunities to recognise the relationship between conceptual diagnoses and real-time symptoms among diseases. These approaches are advantageous for knowledge discovery that could help to achieve a high performance of developing applications for clinical areas.

Originally, scientific literature played an essential role in upgrading the quality of system development. This fact leads to the point that many researchers focus on using a knowledge base in their study. For example, Wang et al. [2017b] used a knowledge resource to develop a system, which was automated to generate disease-pertinent concepts. Huang et al. [2017a] combined multiple populous knowledge sources with building a knowledge graph for helping practitioners to explore realistic clinical queries.

Moreover, Xu et al. [2014b] built their model to enable the system to have a deeper understanding of disease etiology, a model that automatically discovers other patterns specifying semantically similar relationships among diseases. Scientifically, a total of 34,448 disease pairs have been discovered from 21,354,075 MEDLINE records. In contrast, Liu et al. [2009] discovered 3,159 diseases using MeSH annotation of MEDLINE articles to get a comprehensive list of connections between disease and environmental factors. To extract drug or disease symptoms from the knowledge base of MEDLINE, Zeng and Cimino [1998] and Chen et al. [2008] combined text mining and

statistical techniques for automatic acquisition of knowledge from medical domains to identify drug-disease associations. Xu and Wang [2013a] extracted 34,305 unique drug-disease pairs by developing a pattern-based biomedical relationship extraction method from MEDLINE, which was abstracted and compared to 56 602 cancer drug–SE pairs extracted by Xu and Wang [2013b].

2.3.3 Integrating Data and Knowledge

Shah et al. [2019] argued that the combination between the clinical concepts and clinical notes can give researchers evidence in treatment as well as help practitioners in making decisions. A large number of researchers have demonstrated that biomedical literature plays an essential rule in healthcare. Some researchers have taken advantage of biomedical literature in improving the quality of healthcare as well as discovering a broad meaning of concepts related to human disease. Hanauer et al. [2014] and Kavuluru et al. [2013] differentiated between clinical datasets and biomedical literature to discover more fully the relationship and the meanings among concepts. These results brought useful data for further research in the medical domain.

Similarly, Anupindi and Srinivasan [2017] used MESH and ICD based on disease co-occurrence associations to connect the biomedical literature of MEDLINE and the patient data of the phenotypic disease network (PDN), which was created from Medicare Claims for hospitalisations (Hidalgo et al. [2009]). These diseases are generated by differentiating between MEDLINE and patient data to bring more benefits to build innovated models in clinical diagnosis. Escudié et al. [2017] combined the electronic health record

(EHR) and biomedical literature to identify a specific disease. Using text from 741 patients, they obtained 79.3% of the mapped concepts for detected celiac disease. Zhao and Weng [2011] used the linkage of EHR and MEDLINE to build a weighted Bayesian Network Inference model for predicting pancreatic cancer based on selecting 20 common risk factors. The suggestion has a significant accuracy improvement compared to existing representative methods for the prediction of pancreatic cancer.

2.4 Learning Multi-label Classification Models

2.4.1 Problem Transformation

Due to the complexity of multi-label classification problem with respect to the big space of label set, a large number of previous studies have decided to transform the multi-label classification problem as a binary classification and multi-class classification techniques to deal with this issue. The binary classification approach is used to facilitate processing rather than generating multi-label; however, the disadvantages of the method are that, it does not consider label correlations. Similarly, multi-class classification can help to explore label dependencies, but it makes a more prominent space on the original label set.

To exploit the dependencies among labels, Zhang and Zhang [2010] introduced a method that used Bayesian networks to separate the multi-label problem into several single-label classification sub-problems. The conditional dependencies of the labels are based on setting a common parent of labels as additional features for the classifier. The effectiveness of the approach is highly competitive compared well-established methods. A new

chaining approach has been suggested for multi-label classification through binary relevance in (Read et al. [2011]). The method addresses both the disadvantages of the binary method, including the imbalanced data and the dependency between the labels (Dembczyński et al. [2012]) and also maintains acceptable computational complexities. For large datasets, the method achieves superior competitive performance to state-of-the-art methods. Another way to deal with binary relevance for improving the predictive performance of a classifier is to distinguish between the different types of label dependence, including conditional and marginal dependencies (Alvares-Cherman et al. [2012]). In addition, the approach separates the task of multi-label classification into three classes of sub-problem: the estimation of the joint conditional distribution, the minimisation of single-label loss functions, and the minimisation of multi-label loss functions.

Similarly, Tsoumakas et al. [2010] presented a new method that transformed the multi-label problem into several multi-class problems and addressed learning as a single-label classification task. A random subset of the set label was used for building a classifier. The technique can stochastically explore the disjoint and overlapping classes from the construction of the label set. The approach obtained a substantial improvement compared to the original transformed method. Zhang et al. [2015] has proposed a method for promoting an independent rather than a dependent binary relevance for solving the multi-label classification problem. The study used chaining and stacking techniques through pruning methods to deal with the problem. Degree of correlation was estimated among labels for removing unimportant labels. The approach helped reduce the computational costs and raised

predictive performance for the dependent binary relevance model. In addition, Montañes et al. [2014] used chaining and stacking techniques to learn multi-label classification. However, their study concentrated on the type of training data used for model construction and the underlying dependency structure.

2.4.2 Algorithm Adaptation

In suggesting a new algorithm to solve the multi-label problem, Kumar et al. [2013] presented a new technique called the classical method of beam search to address multi-label learning through probabilistic classifier chains. The method helps to deal with the ordering of the tags while training as well as the test time inference. Experimental results of the method yields a state-of-the-art method for learning the multi-label classification. Alali and Kubat [2015] proposed a new classifier-stacking technique to improve the original binary relevance for resolving the problem of multi-label classification. The method helps deal with error-propagation (Senge et al. [2014]) of unknown values of attributes during prediction. Also, the approach can support the removal of unnecessary label dependencies because of classifications which may not be improved by using all class relations. Empirical evidence of the technique outperforms other methods such as dependent binary relevance or Confident Stacking. Similarly, Liu and Cao [2015] proposed a promoted method for multi-label k nearest neighbor algorithm to solve the multi-label problem based on a lazy learning approach. The method helps to exploit more the label relevance through the coupled similarity between class labels. Consequently, the performance of prediction shows a notable outcome compared to multi-label k nearest neighbor.

2.4.3 Ensemble Methods

Multi-label learning was developed extensively in the last decade by a large number of previous works. Most of the researchers emphasise the challenges of multi-label classification problems concerning imbalanced data of the training sets and correlation among labels. Therefore, they focus on dealing with these issues to improve the performance of prediction. The ensemble approach is one of the most effective methods for solving the multi-label classification problem. Tahir et al. [2012] used a heterogeneous ensemble technique for dealing with imbalanced data and labelled relevance to improve the performance of multi-label learning. Experimental outcomes based on six datasets showed that multi-label learners by heterogeneous ensembles could help to overcome over-fitting problems. The approach achieves a high performance results compared to other methods. Li et al. [2013] emphasised that the traditional pairwise constraints among labels play an essential role in multi-label learning. Therefore, the study proposed a new multi-label classification framework for a multi-label ensemble called a variable pairwise constraint projection. The study first used the variable pairwise constraint projection to maintain the correlations between samples and labels. A boosting-like strategy was then adopted to expand the base classifiers. By using ensemble approaches, Mahdavi-Shahri et al. [2016] and Li et al. [2017] introduced a new method to improve the performance of multi-label learning. These approaches focused on dealing with the imbalanced data, which helped to exploit dependencies among different labels.

Recently, heterogeneous information networks have increased in popularity because they can help to explore more hidden knowledge regarding relationships among different types of entities for both data samples and

class labels. Kong et al. [2013] took advantage of heterogeneous information networks to extract multiple types of relationships among different classes of labels by exploiting the linkage structure. These relationships help to facilitate the multi-label classification process. Using heterogeneous information networks for multi-label learning is an effective approach to solve label correlations because the correlations among different class labels may be difficult to learn directly while some may be missing. Empirical results indicate that the performance of multi-label classification is boosted by using heterogeneous information networks. Similarly, Zhou and Liu [2014] presented an activity-edge centric multi-label classification framework based on heterogeneous information networks. The study first combined structures used between the primary social network and multiple associated activity networks to generate a unified multigraph based on edge classification. The study secondly used the structure affinity and the label vicinity to create a unified classifier based on multiple activity networks. Then, an algorithm was proposed to refine the classification result by different activity-based edge classification schemes from multiple activity graphs. The experimental result showed that the approach achieved a high performance compared to existing methods.

2.5 Summary

The thesis endeavoured to discover the knowledge from the medical domain in developing classification models with both accuracy and reliability. This chapter, therefore, examines various domains of mining data in addressing classification problems. It is clear that machine learning plays an important role in the discovery of new knowledge in different domains

(David [2007], Koh et al. [2011]). In particular, it has a significant effect on improving the performance of decision support systems in healthcare (Herland et al. [2014], Rosset et al. [2010], Holzinger [2016], Yoo et al. [2012]). There are many approaches to solve the classification problems. In the review of the chapter, it can be asserted that many studies have investigated binary classification problems and developed related data mining techniques. Table 2.1 shows the brief information about learning binary classification. Health status measurement prognostication is known as a first approach for supporting health practitioners in improving health knowledge and clinical decisions (Miyazaki [2013], Prakash et al. [2006], Wagner and Draper [1984], Keegan et al. [2012]). By focusing on the prediction of health risks, many researcher have developed a binary classification to help the health risk prediction (Yeh et al. [2011], Neuvirth et al. [2011], Nguyen et al. [2014], Yang and Loog [2016]). However, these model often use a limitation of input data which lead to uncertain results. To improve these challenges, graph-based approaches have been used by (Guillory and Bilmes [2009], Sun et al. [2009], Ji et al. [2010], Luo et al. [2014], Hwang and Kuang [2010], Chen et al. [2016a]) to deal with the types of disease through traditional diagnosis. Using a graph to discover the real dataset can help to improve the accuracy of the classification models. Although the graph could bring more benefits to improve the accuracy of the classification models, the way to build these graphs still have a limitation. Therefore, this thesis is expected to provide a new method to build a graph which helps to boost the performance of the classification models.

As an advantage of the knowledge graph, ming data under concepts and

TABLE 2.1: The weakness of related work for health risk prediction by binary classification

Category	**Author**	**The gaps of the existing research**
Health Status Measurement	Miyazaki [2013]; Prakash et al. [2006]; Wagner and Draper [1984]; Keegan et al. [2012]	These scoring systems just provide basic information to support health practitioners in improving health knowledge and clinical decisions.
Health Risk Prediction	Yeh et al. [2011]; Neuvirth et al. [2011]; Nguyen et al. [2014]; Yang and Loog [2016]	By analysing the real dataset, a few of the binary classification models is developed to predict the health risk status. These model have a significant impact on support practitioners. However, these models have limited on discovering the real dataset.
Classification by Mining Heterogeneous Graphs	Guillory and Bilmes [2009];Sun et al. [2009];Ji et al. [2010];Luo et al. [2014]; Hwang and Kuang [2010]; Chen et al. [2016a]	By using a heterogeneous information graph, researchers may have deeply discovered elements of the real dataset. However, these methods of building a graph still have challenges because of the complex of the data.

relationship among concepts has contributed significant benefits in discovering hidden knowledge. A large number of researchers have shown that mining the data by using concepts can bring high performances of discovering knowledge (Lee et al. [2006], Gao et al. [2017], Ni et al. [2016], Choi et al. [2014], Wang and Akella [2015], Karpagam et al. [2016]). Knowledge graphs have been successfully considered for using in data mining by researchers (Xu et al. [2014a], Bordes et al. [2011], Nguyen et al. [2017], Shi et al., Voskarides et al. [2015]). These works have motivated the research by proposing a framework to build a knowledge graph, which is based on MEDLINE to improve the performance of exploring knowledge in the medical domain. Applying the knowledge graph can help increase the accuracy of classification models.

Although using a knowledge graph has more benefits in discovering knowledged, it is a challenge to combine different data sources for building one. Combining different sources to create a cross-mapping can help to discover more semantic relationships among terminologies. The cross-mapping of different medical sources is completed by a large number of researchers (Pereira et al. [2006], Soualmia et al. [2013], Srinivasan et al. [2002], Schriml et al. [2018], Saitwal et al. [2012], Zhang et al. [2016]). By using the advantages of combining sources in medical domain, numerous researchers have developed applications to enhance the quality of clinical diagnosis (Shah et al. [2019], Hanauer et al. [2014], Kavuluru et al. [2013], Anupindi and Srinivasan [2017], Hidalgo et al. [2009], Escudié et al. [2017], Zhao and Weng [2011]). In the thesis, an approach combining the observational data and biomedical literature has been conducted for building a heterogeneous information graph, which can help to boost the performance

of classification models.

TABLE 2.2: The weakness of related work for health risk prediction by multi-label classification

Category	**Author**	**The gaps of the existing research**
Problem Transformation	Zhang and Zhang [2010], Read et al. [2011], Tsoumakas et al. [2010], Zhang et al. [2015]	Transforming the multi-label classification problem as a binary classification and multi-class classification techniques to facilitate learning has succeeded by many researchers. However, this method has limited on solving label correlations, which play an important role in learning multi-label classification.
Algorithm Adaptation	Kumar et al. [2013], Alali and Kubat [2015], Liu and Cao [2015]	This method could help increase the accuracy of classification by their optimised algorithm. This approach has limited on dealing with imbalanced data of the training sets and the obstacle of determining the correlation among labels.
Ensemble Methods	Tahir et al. [2012],Li et al. [2013], Mahdavi-Shahri et al. [2016], Li et al. [2017], Kong et al. [2013], Zhou and Liu [2014]	This is one of the methods that has been used by many researchers recently for dealing with the multi-label classification, especially, applying a graph in learning multi-label classification. However, this method leads to an increase in the space of label correlations.

Besides solving the binary classification, the problem of multi-label classification is addressed in the thesis. Table 2.2 presents a common review of

three different approaches to deal with the multi-label classification. Handling multi-label classifications as well as predicting multiple diseases remains a challenging task for many researchers due to its complexity. To determine whether an example has multiple labels, a multi-label classification model requires multiple processing stages. To deal with the complexity of a multi-label learning problem, researchers often break down a big problem into small problems to facilitate learning or focus on one particular issue to increase the accuracy of classification. Problem transformation is one of the options that is used to dealt with the multi-label problem. Based on the approach, the multi-label classification problem is transformed into several single-label classification problems (Zhang and Zhang [2010], Alvares-Cherman et al. [2012]) or several multi-class problems (Tsoumakas et al. [2010], Zhang et al. [2015], Montañes et al. [2014]), that superseded multiple techniques in the past. The approach of problem transformation facilitates the learning of multi-label classification.

Meanwhile, some other researchers have chosen to focus on proposing an optimal algorithm to improve classification accuracy (Kumar et al. [2013], Alali and Kubat [2015], Liu and Cao [2015]). All these methods show that the essential challenges of learning multi-label classification are due to imbalanced data of the training sets and the obstacle of determining the correlation among labels. Therefore, some researchers recently applied the ensemble method (Tahir et al. [2012], Li et al. [2013]) to deal with these issues. Using heterogeneous information networks (Kong et al. [2013], Zhou and Liu [2014]) to deal with the multi-label classification problem is one of the more efficient methods to discover the relevance label. The approach is able to exploit multiple types of relationships among different class labels

based on discovering the linkage structure of the heterogeneous information graph.

Although a large number of researchers have succeeded in addressing the multi-label problem, however, exploiting correlations between labels still has limitations, especially in the field of multiple disease diagnosis, a model that needs to deal with cause and effect relationships. These relationships are extremely important because one symptom can be a causal effector of many diseases; some symptoms may be causal to a particular disease. Therefore, the thesis represents the characteristics of a patient under a graph to discover relationships among underlying patterns of a dataset. Then the graph is separated into two different domains to boost leaning the correlation between diseases. In particular, the integration of evidence medicine into classification models is used to help increase the reliability of that model in practical applications. Moreover, it also helps in finding the profound relationships between diseases as well as their corresponding symptoms.

Chapter 3

Binary Classification for Health Risk Prediction using a Heterogenous Information Graph

In this Chapter, the study proposes an innovative classification model for knowledge discovery from patients' personal health repositories. The model discovers medical domain knowledge from the massive data in the National Health and Nutrition Examination Survey (NHANES) based on a heterogeneous information graph. The graph is built using combining *Pearson Correlations* and *Semantic Relations*. On the basis of the model, an innovative method is developed to help uncover potential diseases suffered by people and to classify patients' health risk. The proposed model is evaluated by comparing it to a baseline model, which was also built on the NHANES data set in an empirical experiment. The study makes significant contributions to the advancement of knowledge in data mining with an innovative classification model, specifically crafted for domain-based data. In addition, by accessing the patterns of various observations, the research contributes to the work of practitioners by providing a multifaceted understanding of

individual and public health.

3.1 Introduction

Data mining is used in the healthcare area in order to assess the effectiveness of treatment. In recent years, many researchers have spent time and effort on assessment of the risk of diseases. The research mainly focuses on supporting physicians and other medical practitioners in making secure and effective clinical decisions relating to their patients. Along with physicians' experience, medical databases are considered to be valuable sources to provide evidence to facilitate effective decision making. Risk assessment has been undertaken to improve the assessment and management of diseases by (Cheng et al. [2017], Chin et al. [2015]). These approaches have helped to improve accuracy as well as to reduce the risks relating to disease treatment. However, the diversity and complexity of terms and concepts in the documents available for study may limit the accessibility of information that is effectively retrieved.

Semantic similarity is widely used in improving the performance of mining data to retrieve the highest amount of data. Many researchers have taken advantage of Semantic similarity to achieve a significant improvement in understanding of the issues involved. Ni et al. [2016] proposed a new method to measure similarity among concepts. The approach represents concepts as continuous vectors which are utilized to accumulate similar pairwise arrangements among pairs of concepts. The study improves the measurement of semantics among documents. Similarly, to guarantee the semantic measure of similarity between the medical terms, Karpagam et al. [2016] suggested a new method which combines the sources of disease

concepts and biomedical resources to identify the medical terms for automatically extending the Disease Ontology. The trend of the semantic similarity for mining data has increased recently because semantic similarity has contributed a significant improvement in discovering new knowledge.

The increasing volume of datasets is helpful for medical professionals who wish to improve the quality of healthcare. In relation to medical diagnosis, some researchers have developed predictive models for the classification of clinical risks (Kim et al. [2014], Sabibullah et al. [2013]) and for predicting diseases (Chang et al. [2011], Huang et al. [2012]). However, there are still issues relating to applying real datasets in mining. Most of the datasets lack data labels. Moreover, the lack of features for representing all types of networked data may have a negative effect on discovering knowledge. As a result, mining these data may not bring about a high performance result because of the complexity of the real dataset. The datasets also have various types of objects. Therefore, presenting these datasets by using a new method is necessary to advance effective mining.

Due to the limitation of the real datasets, many researchers use a heterogeneous graph to represent the data for developing classification models. Ji et al. [2010] proposed a new algorithm to predict the label for each object by separating the different types of links and objects which can be applied on the heterogeneous graph. This approach shows a significant improvement in the task of surmounting the classification problem. Following some successful trials, there has been an increase in the number of works on classification in the heterogeneous graph (Sun et al. [2009], Luo et al. [2014], Wan et al. [2015], Kong et al. [2012]). This approach has helped the model to be more meaningful compared to mining using traditional methods. In

order to diagnose disease, Chen et al. [2016a] proposed an algorithm called a semi-supervised heterogeneous graph on health (SHG-Health) to predict the risk of mortality and morbidity of patients based on the health examination data. The model is built on mining the heterogeneous graph only based on the neighbouring nodes' information. This work has a significant result in discovering the different types of relationship in the heterogeneous graph for health risk prediction by classification.

3.2 Research Problem Formulation

This study focuses on the observation data to build a binary classification for predicting the health risk status. The following definitions illustrate the research problem.

Definition 3. 1. ***[Electronic Health Records]***

The Electronic Health Records are a 3-tuple $\mathbb{R} = \{\mathcal{P}, \mathcal{A}, \mathcal{M}_{\mathcal{P}}^{\mathcal{A}}\}$*, where*

- $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$ *is the set of patients in the entire dataset and* $|\mathcal{P}| = m$*;*
- $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ *is a set of attributes and* $|\mathcal{A}| = n$*. Each attribute has a label label(a) that marks the semantic meaning of a;*
- $\mathcal{M}_{\mathcal{P}}^{\mathcal{A}}$ *is a matrix constructed by* $\mathcal{P} \times \mathcal{A}$ *with values taken from a survey with questions defined by* $\mathcal{A}$ *of patients* $\mathcal{P}$*.* □

Definition 3. 2. ***[Patient Health Profile]***

The health profile $\mathcal{HP}(p)$ *of a patient* $p \in \mathcal{P}$ *is defined as a vector* $\overrightarrow{p} = \{\langle a_1,$

$w_1\rangle, \langle a_2, w_2\rangle, \ldots, \langle a_n, w_n\rangle\}$, *where* $a \in \mathcal{A}$ *and* w *is the value of attribute* a *on patient* p. □

Definition 3. 3. ***[Research Problem]***

Let $P = \{p_i, i = 1, \ldots, \imath\}$ *be a set of patients and* $P \in \mathcal{P}$; $\mathcal{S} = \{s_1, \ldots, s_K\}$ *be a set of classes, where each* s *is a disease and* K *is the number of classes. Given a training set of patients* $P_t = \{p_j, j = \imath + 1, \ldots, m\}$ *and their respective health profiles* $\mathcal{HP}(p_j)$, *with* $y_j^k = \{0, 1\}, k = 1, \ldots, K$ *is provided for describing the likelihood of* p_j *belonging to a class* s_k, *the research problem is to learn a binary prediction function* $f(y^k|p)$ *and use it to classify* $p_i \in P$ *into* $\{s_k\} \subset \mathcal{S}$ *for prediction of the patients' health status in terms of the set of diseases defined in* $\mathcal{S}$. □

3.3 Framework

3.3.1 Data Correlation

Pearson correlation is an approach that can help to identify the secure connection between two factors and can help machine learning to obtain an optimum result in data mining. In healthcare data, some researchers have used the *Pearson correlation* coefficient to investigate the relationship between diseases. Ha et al. [2017] used the *Pearson correlation* coefficient to identify the relationship among diseases for predicting the prognosis of high-risk patients. The *Pearson correlation* is used widely in the medical domain to analyse data by Zhou et al. [2014], Tsanas et al. [2013], Torres et al. [2012] for discovering the underlying patterns in the real dataset. All of the works demonstrate the critical effect of the *Pearson correlation* coefficient in mining data.

In this study, the *Pearson correlation* is adopted to identify the potential connection of different types of data in the dataset. The *Pearson correlation* value indicates the strength of relationship between data. Each data type is transformed to a node in graphic format. The connection between those data evolves into edges linking the nodes. The value assigned to the edges represents the strength of the links. By using this approach, a heterogeneous information graph (HIG) can be constructed. The HIG can be used to help discover patterns underlying the data.

To identify the links between data and measure their strengths, the following *Pearson correlation* formula is exploited Egghe and Leydesdorff [2009].

$$\rho = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{3.1}$$

where x and y are two random elements in $\mathcal{A}$. The values of the *Pearson correlation* coefficient ρ reveals how one data affects the others and distinguishes the correlation between different data. With the values, data with strong connection are clustered in a common class.

As defined in **Electronic Health Records**, $\mathcal{M}_{\mathcal{P}}^{\mathcal{A}}$ is a matrix constructed by $\mathcal{P} \times \mathcal{A}$, Therefore, given an attribute $a \in \mathcal{A}$, using a mapping function that returns all a's corresponding values in patients' health profiles can be defined:

$$\Omega(a_i) = \{w_i | \langle a_i, w_i \rangle \in \vec{p}, \forall p \in \mathcal{P}\} \tag{3.2}$$

Let $\mathcal{C}_{\mathbb{R}}$ be a set of knowledge specifying the class revealed by correlation

of the data in $\mathbb{R}$, the Electronic Health Records; c be a concept in the health domain and $c \in \mathcal{C}_{\mathbb{R}}$; γ be a threshold determining whether or not two attributes are strongly related. **Algorithm 3.1** presents how the knowledge is discovered from the healthcare dataset.

Algorithm 3.1 Knowledge Discovery from Health Data.

INPUT: $\mathbb{R} = \{\mathcal{P}, \mathcal{A}, \mathcal{M}_{\mathcal{P}}^{\mathcal{A}}\}$;
OUTPUT: $\mathcal{C}_{\mathbb{R}}$;
Let $\mathrm{K}_{\mathbb{R}} = \emptyset, isInc = false$;
for each $a_i \in \mathcal{A}$ **do**
  $w_i \leftarrow \Omega(a_i)$;
  for each $a_j \in \mathcal{A}, a_i \neq a_j$ **do**
    $w_j \leftarrow \Omega(a_j)$;
    $\rho_{i,j} \leftarrow pearsonCorrelation(w_i, w_j)$;
    if $\rho_{i,j} \geq \gamma$ **then**
      for each $c \in \mathcal{C}_{\mathbb{R}}$ **do**
        if $(a_i \in c) \wedge (a_i \notin c) \wedge (isInc = false)$ **then**
          $c = c \bigcup \{a_j\}$;
          $isInc = true$;
        else
          if $(a_j \in c) \wedge (a_j \notin c) \wedge (isInc = false)$ **then**
            $c = c \bigcup \{a_i\}$;
            $isInc = true$;
          end if
        end if
      end for
      if $isInc = false$ **then**
        $c = \{a_i, a_j\}$;
        $\mathcal{C}_{\mathbb{R}} = \mathrm{K}_{\mathbb{R}} \bigcup \{c\}$;
      end if
      $isInc = false$;
    end if
  end for
end for
return $\mathcal{C}_{\mathbb{R}}$

Algorithm 3.1 fist uses a loop to check all attributes of the dataset. Then, the algorithm uses a mapping function by Eq.(3.2) to return all corresponding values for each attribute a_i. Next, another for loop is used to check

the rest of attributes a_j, $a_i \neq a_j$. Then, the *Person Correlation* coefficient ρ between a_i and a_j is conducted by Eq.(3.1). For all ρ identified great γ, a loop is used to check whether or not concepts belong to the specific knowledge $\mathcal{C}_{\mathbb{R}}$. If each concept learned between a_i and a_j belongs to the health knowledge domain and concepts do not exist in $\mathcal{C}_{\mathbb{R}}$, these concepts learned between a_i and a_j are added into $\mathcal{C}_{\mathbb{R}}$. Repeating this process for the rest of attributes reveals all possible concepts for $\mathcal{C}_{\mathbb{R}}$. This algorithm uses three nested loops for learning concepts for $\mathcal{C}_{\mathbb{R}}$. The complexity of the algorithm is identified by $O(n^3)$.

The health knowledge, $\mathcal{C}$, discovered by mining health data $\mathbb{R}$, is a set of health concepts as defined below:

Definition 3. 4. ***[Latent health knowledge]***

Latent health knowledge, denoted by $\mathcal{C}$, is a set of health concepts, in which each element is $c := \langle \mathcal{M}_A^A, \vec{\mathcal{M}}_A^A \rangle \in \mathcal{C}$, where $\mathcal{M}_A^A$ is a matric $A \times A, A \subset \mathcal{A}$ and $\mathcal{A} \in \mathbb{R}$. For each pair $(a_i, a_j) \in \mathcal{M}_A^A$, the value of weight $(a_i, a_j) \in \vec{\mathcal{M}}_A^A$ indicates the strength level of correlation between a_i and a_j. □

Table 3.1 presents a couple of health concepts discovered by **Algorithm 3.1**. A concept comprises attributes a_1 (MCQ160D), a_2 (MCQ160B), a_3 (MCQ160C) and a_4 (MCQ160E). They are strongly connected to one another and thus clustered in a common concept. (In fact, these attributes are commonly related to heart issues.) Another concept consist of a_6 (LBXBPB) and a_7 (LBDBPBSI). These attributes also have a strong relationship and have been clustered - they are actually all about blood. Attribute a_5 (WTSH2YR) however, has no relationship with others listed on the table and are excluded from the "heart" and "blood" concepts. (They may belong to other classes that are not shown on the table due to the limit of space here).

TABLE 3.1: The Pearson Correlation coefficient. The code of MCQ160D (a_1), MCQ160B (a_2), MCQ160C (a_3) and MCQ160E (a_4) belong to the Heart class, where LBXBPB (a_6) and LBDBPBSI (a_7) belong to the Blood class and WTSH2YR (a_5) belongs to other class, where the emphasised values indicate strong connection of attributes

	a_1	a_2	a_3	a_4	a_5	a_6	a_7
a_1	**1**	**0.264**	**0.415**	**0.321**	-0.032	0.039	0.041
a_2	**0.264**	**1**	**0.292**	**0.401**	-0.019	0.004	0.005
a_3	**0.415**	**0.292**	**1**	**0.416**	-0.020	0.047	0.049
a_4	**0.320**	**0.401**	**0.416**	**1**	-0.049	-0.001	0
a_5	-0.032	-0.019	-0.020	-0.049	**1**	-0.054	-0.066
a_6	0.039	0.004	0.047	-0.001	-.0.054	**1**	**0.981**
a_7	0.041	0.005	0.049	0	-0.066	**0.981**	**1**

3.3.2 Semantic Relations within Data

Domain knowledge has been widely used in data mining to help improve the performance of systems in a specific domain. Such relationships can be generated for a range of semantic relations as well as syntactic relations. Xu et al. [2014b] constructed a model to enable the system to have a deeper understanding of diseases, a model that automatically discovers other patterns specifying semantically similar relationships among diseases. By presenting the concept detected from a document into a graph, Ni et al. [2016] proposed a new method to measure similarity among concepts. Then, the approach represents concepts as continuous vectors which are utilized to accumulate pairwise similarity among pairs of concepts. This study demonstrated the improvement in measuring semantics among documents.

Discovering knowledge from the medical domain has been successful by analysing data for generating a semantic network (Abacha and Zweigenbaum [2011]). A semantic network is set up by linking all of the concepts

(the semantic class, or the node with the same type). There are a large number of relationships between two semantic types being set up from the semantic network. These relationships help to provide more chances to improve the performance of decision support systems in healthcare.

In this study, to conduct a risk prediction for each patient based on the heterogeneous graph, the research considers each type of concept node as a semantic class. Then the research builds up a model to predict the risk of each disease that belongs to the different class by comparing the effect among these classes. In this study, the same type of nodes are set up by considering all attributes with the same semantics for constructing the graph. All observation data in the NHANES dataset are categorised into eight classes. The eight medical semantic classes are presented in Table 3.2 with a brief description. Each class consists of a group of diseases. For example, the Hepatitis class contains both Hepatitis B and Hepatitis C. By adopting semantic knowledge in our work, underlying patterns from the data can be discovered.

The domain health knowledge, $\mathcal{S}$, acquired by categorizing the attribute label in the health dataset $\mathbb{R}$, is a set of semantic health concepts, which is defined as:

Definition 3. 5. ***[Domain Health Knowledge]***

Domain Health Knowledge is a set of health concepts, $s \in \mathcal{S}$*, where* $s = \{lebel(a_1), lebel(a_2), ..., lebel(a_n)\}$*,* $a \in \mathcal{A}$*, which is a set of attribute labels connected by the semantic relation of "related-to", and* $\mathcal{A} \in \mathbb{R}$*.* □

TABLE 3.2: Semantic categories

Class	**Description**
Kidney Conditions	All the attributes related to kidney disease.
Hepatitis	All types of hepatitis such as A, B, and C. In addition, some questions are asked related to hepatitis, for example, "Have you ever received Hepatitis A vaccine".
Diabetes	Urine or blood lab test.
Blood Pressure and Cholesterol	All the lab tests relating to blood.
Heart disease	Questions such as "Has a doctor ever told you that you had a heart attack, coronary heart disease, or congestive heart failure?". In addition, the doctor may ask about angina (angina pectoris).
Respiratory Disease	Attributes of respiratory disease, e.g., asthma, emphysema, thyroid problem, chronic bronchitis.
Profile	Personal demographics such as age, weight, and gender.
Others	Miscellaneous attributes or attributes where ground truth can not be obtained.

3.3.3 Heterogeneous Information Graph Construction

In this study, a heterogeneous graph is constructed consisting of the different types of nodes based on health examination records. By presenting the data as a heterogeneous information network, hidden information that exists in the real dataset can be discovered. The National Health and Nutrition Examination Survey (NHANES) dataset is used to build the heterogeneous information graph. The NHANES dataset consists of 2585 attributes, covering a wide range of information about an individual, such as personal demographics, observations, laboratory tests, and diagnostic reports. After careful examination, these characteristics were manually categorised into eight classes based on their intuitive semantic relations. *Hepatitis*, *Kidney conditions*, *Diabetes*, *Heart disease*, *Blood Pressure and Cholesterol*, *Profile Respiratory Disease* and *Others*, are presented in Section 3.3.2

As mentioned above, this study considers using a new approach for setting up links between two nodes of the heterogeneous information graph. Considering each attribute in the NHANES dataset as a data object, the underlying links connecting objects are discovered by adopting the *Pearson correlation* coefficient, which is a powerful technique to measure the linear correlation between two variables. Denote a *Pearson Correlation* coefficient value by ρ, two objects (two attributes) v_1 and v_2 hold valid connection if $\rho(v_1, v_2) \geq \gamma$, where γ is a threshold defining the validity of object connection. The range of the coefficient value is set between -1 and 1. The relationship between the two objects is perfectly negatively linearly related if the coefficient value is approximately -1. In contrast, the relationship between two nodes is more strongly correlated if the coefficient value is in the possible range of nearly 1. Apparently, a larger ρ value indicates a stronger

connection between the two objects.

A heterogeneous information graph is then constructed with the *semantic classes* and the link connecting objects is defined by the *Pearson correlation* coefficient. A heterogeneous information graph (Ji et al. [2011a], Sun et al. [2009]) is adopted to help explore all different types of data since the dataset has many types of categories. The graph can cover different types of nodes which have multi-typed relationships of data items being linked. In the design of our model, the unhealthy status is assessed by a classification model, using a heterogeneous information graph constructed by the data collected from observation of a patient.

Definition 3. 6. ***[Health Knowledge Graph]***

The Health Knowledge Graph graph is a 2-tuple, $G :=< V, E >$ *with an object mapping function* $\varphi : V \rightarrow A$ *and a link type mapping function* $\psi : E \rightarrow R$*, where*

- V *is a set of vertices, in which each element* v *is an attribute* $a \in \mathcal{A}$ *and* $A : \varphi(v) \in A$*;*
- E *is a set of edges, in which each element is a semantic relation* r *in the relation type set* $R : \phi(e) \in R$*, where* $R = \mathcal{C} \bigcup \mathcal{S}$*.* □

The coefficient value of the *Pearson correlation* (ρ) is an essential factor in building the heterogeneous graph. It is calculated based on the influence between two attributes (variables) belonging to the different objects. The coefficient values of the *Pearson correlation* is calculated through the use of K attributes. However, some of the coefficient values may have a minus value. Nevertheless, the coefficient value of some (ρ) are equal 0 because they do not have an influence or relation between two attributes. The *Pearson correlation* (ρ) needs to have the threshold value of γ for identifying a validity link between two objects.

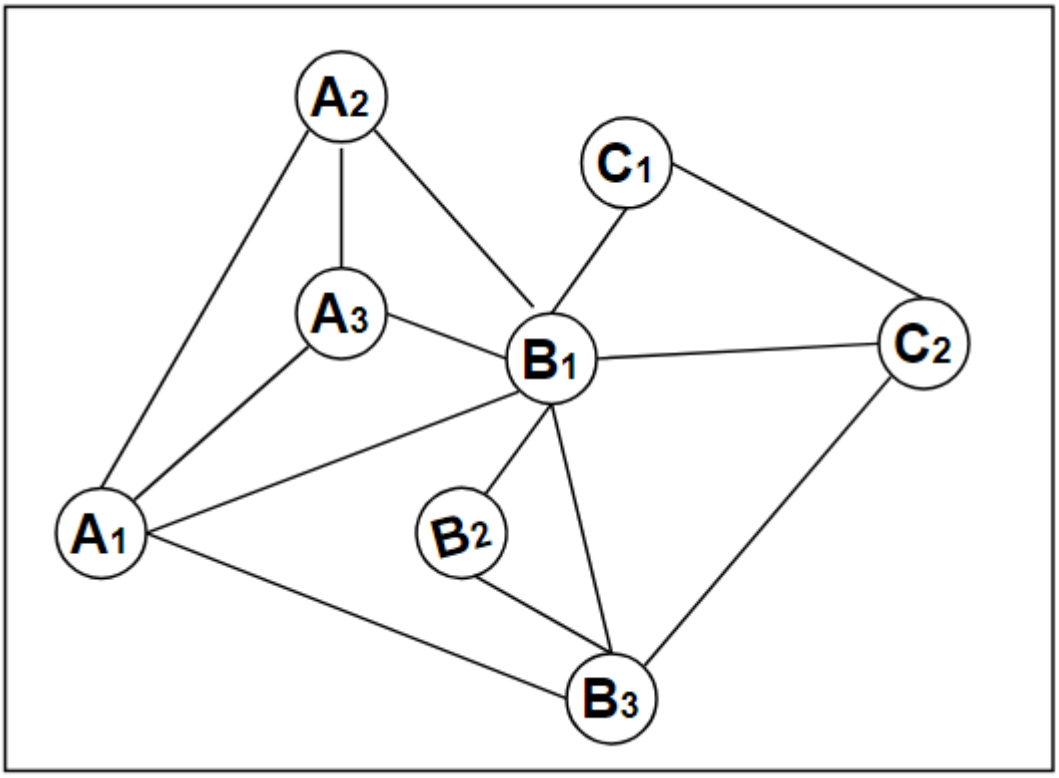


FIGURE 3.1: A sample heterogeneous information graph

Figure 3.1 illustrates a subgraph of the heterogeneous information graph constructed using the NHANES dataset. Three types of objects are illustrated, where *A* is a class for *Heart Disease*, *B* for *Patient Profile* and *C* for *Kidney Condition*. As we can see, the profile of aged (B1) is related to most of the nodes because people of all ages can get any disease. However, height (B2) can only link to B1 and B3 (weight) because height does not affect kidney or heart disease. However, B3 plays an important aspect in causing a heart attack (A1) as well as a diseased kidney (C2). There is no link between heart disease and kidney conditions. Among aspects in *Heart Disease* and *Kidney Conditions*, links always exist because these aspects are in the same class.

3.3.4 Binary Classification Model

With the availability of the heterogeneous information graph, a function can be learned from the training data that formalises the profile of a patient for her state of healthiness or unhealthiness regarding a disease x:

$$f(x) = \sum_{i=1}^{k} v_i \times \rho(x, v_i) \times \alpha + \sum_{j=1}^{k} v_j \times \rho(x, v_i) \times \beta \tag{3.3}$$

where $\varphi(x) = \varphi(v_i)$ and $\varphi(x) \neq \varphi(v_j)$. Each element v_i or v_j is an attribute $a \in \mathcal{A}$. $\rho(x, v_i)$ is a *Pearson Correlation* coefficient, which is identified for building a graph for disease x. Also, α and β are two coefficients adopted to clarify the contribution of *latent health knowledge* and *domain health knowledge* in the classification model.

Based on $f(x)$, a patient's health status can be modelled against a single disease, x:

$$y(x) = \begin{cases} 1, & \text{if } f(x) \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

where θ is a threshold to determine the final class, "healthy" or "unhealthy", for the patient. When checking against multiple diseases $x \in \mathcal{X}$, an overall model can then be defined on the basis of *Eq*. 3.4:

$$y(\mathcal{X}) = \Pi_{n=1}^{k} y(x_n), \text{where } x \in \mathcal{X}, |\mathcal{X}| = k \tag{3.5}$$

In this model, there are three important parameters which have a substantial effect on running the training model. First, if we consider the coefficient of the same and different semantic groups α and β, these two parameters show the effect between the different semantic groups, where α+β=1.

If the coefficient of α is nearly equal 1 (β approximates 0), other groups do not have a role in this model. The model favours *domain health knowledge* and omits *latent health knowledge*. The model only considers the factors in the same group to predict the high risk to health for a patient. For example, the model uses attributes of the Hepatitis B class to predict whether these patients belong to the Hepatitis B class (unhealthy) or not. In contrast, if the coefficient of β is nearly equal 1 (α approximates 0), the model only considers attributes of the different semantic group to predict the high-risk to health for patients. The *latent health knowledge* takes place and the *domain health knowledge* is fades. The model uses the attributes that are not related to the Hepatitis B class to predict whether or not these patients belong to the class. When α and β are of the same value (α =0.5 and β = 0.5), *domain health knowledge* and *latent health knowledge* are equally considered in the classification model. Another parameter is the threshold value of θ. The parameter has a direct effect on deciding whether the patient belongs to a healthy or unhealthy case.

An optimisation **Algorithm** 3.2 is also developed to find the best values for variables α, β and θ in the model:

In **Algorithm** 3.2, at the third-step, the first value of the threshold θ is set equal to 0.1. Following a loop is used to set the value of α and β, where α plus β equal to 1. Then calculating the performance of the model is to define the best value of F-measure for function *f(x)*. There were 11 values of F-Measure for the first round with θ being 0.1. Then, the algorithm can define the max value of F-Measure as well as identify the best threshold for α and β. At the four-step, the algorithm uses the value of α and β from the third-step is to set a loop for θ running from 0. to 0.9 and calculate the

Algorithm 3.2 Optimisation Algorithm

1: **INPUT**: $\beta = \{0, 0.1, \ldots, 1\}$, $\alpha = \{0, 0.1, \ldots, 1\}$ and $\theta = \{0.1, \ldots, 0.9\}$
2: **OUTPUT**: The best value of θ, α and β
3: set $\theta = 0.1$;
for each $\alpha \in \{0, 0.1, \ldots, 1\}$ where $\beta \in \{1, 0.9, \ldots, 0\}$

- Calculate weight value for Eq.(3.3)
- Select the best α and β

4: set the best α and β
for $\theta = 0.1, 0.2, ..., 0.9$

- Identifying the performance of Eq.(3.3)
- Select the best θ

5: repeat step 1 for the best θ
6: repeat step 2 for the best α and β
7: repeat step 3 and 4 until convergence
8: return θ, α and β

performance of the model. Then the max value of F-measure for the classification model is defined based on 9 values of F-Measure. Then the best value of θ is identified. Repeating the third-step with a new value of threshold θ is to confirm the best value of α and β. If these values are converged, the four-step will repeat until all values of α, β and θ are converged. The algorithm uses two nested loops for identifying the value of α, β and θ. Therefore, the complexity of the algorithm is $O(n^2)$.

3.4 Empirical Experiments for Evaluation

3.4.1 Dataset

In this study, the NHANES dataset was used for the experiment. The NHANES dataset has been collected since 1960 in the United States. It comprises a series of surveys including a variety of health and nutrition measurements with different population groups. This dataset provides significant health knowledge to help identify and assess the risk of disease for Americans. The dataset covers a wide range of health assessments, such as lab tests, physical examinations and personal habits. The NHANES dataset contains 9770 participants with more than 2585 attributes which are categorized by Table 4.1 except for the characteristics related to food and nutrition. In the experiments, the dataset collected from 2013 to 2014 was used to train and test the classification model.

Due to the missing data, this research only considered the attributes which have a limited level of sparseness. Figure 3.2 shows the percentage of the missing data of the NHANES dataset. Also, data for foods and nutrition are not considered. As a result, there are 318 attributes used in the experiments.

The dataset was manually assessed to generate the ground truth based on 13 attribute correspondents with 13 diseases in Table 3.4, which were selected for experiments in this model. Only patients who were clear of all thirteen diseases were considered healthy. As a result, 5144 out of 9770 participants were identified healthy, and 4626 participants were unhealthy. Table 3.5 illustrates the statistics of the NHANNES dataset. Before applying the dataset to experiments, it was pre-processed first as it contained many

TABLE 3.3: NHANES attributes by categories

Type	Category	Attribute description
Patient Profile	***Demographics***	age, marital status, gender, education level, residential suburb, annual income, weight, people according to age groups, total number of people in the family/household, language used in interview
	Habit	consumption behavior, diet behavior and nutrition, physical activities, smoking, alcohol use, drug use
Questionnaire	***Mental Health***	questions regarding sleep disorders, depression, cognitive problems
	Current Health Status	diabetes, diagnosis of hepatitis B or hepatitis C kidney disease, sexual behaviour, osteoporosis, cardiovascular disease, dermatology, disability, immunization, oral health
	Health Conditions	asthma, childhood and adult, anaemia, psoriasis, heart, diseases, arthritis, blood transfusions
	Family History of Disease	asthma, diabetes, heart attack/angina
Examinations	***Physical***	weight, height, recumbent length, body mass, circumference muscle strength, blood pressure
	External	femur, neck, head circumference, leg, arm
	Other	trochal term, abdominal diameter, teeth, gum disease, oral hygiene, impression of soft tissue condition, denture/partial, denture/plates
Lab Tests	***Biochemical***	albumin, cholesterol, glycol haemoglobin, insulin, glucose, vitamin B12
	Blood	blood metal weights, blood lead, blood cadmium, blood mercury, blood selenium, blood manganese
	Urine	urinary arsenic, urinary creatinine, sugar, iodine, mercury, metal, urine pregnancy, trichomonas
	Other	toxocara, hepatitis, HIV antibody, human papilloma virus, nitrate, thiocyanate

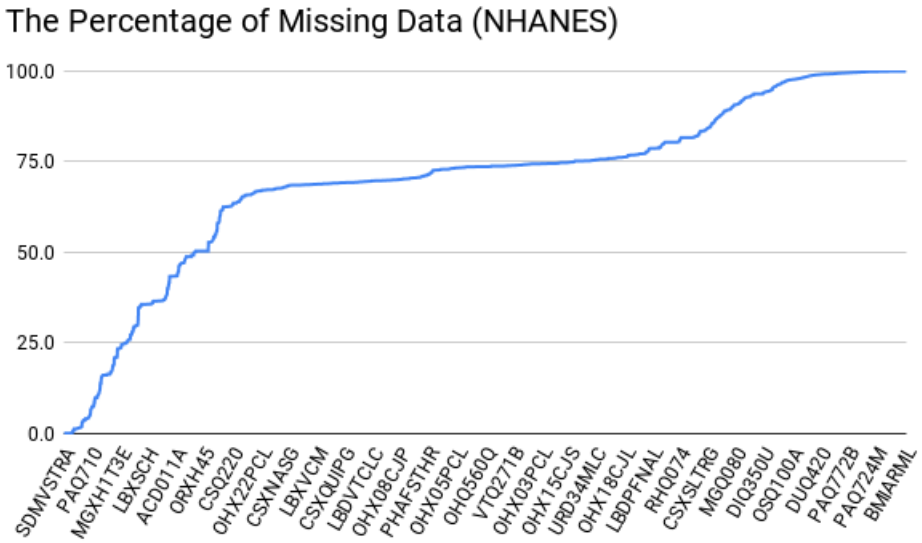


FIGURE 3.2: The percentage of missing data

noisy data. The values of attributes were normalised into a unified form. Then, the missing data were handled and replaced by the average of all values in the respective attribute.

Data Normalisation

The dataset needs to be pre-processed to be transformed into a unique format for the experiment. For example, a nominal data type such as gender is presented as equal to 0, and 1 instead of the format type being male and female. The ordinal data type such as general health condition is presented as 0, 0.5 and 1 instead of presenting as bad, good and excellent. With the data type such as a range of blood test (50 to 150), the research uses the standard of maximum and minimum to transfer data into the range of [0, 1]. Zero is set for a negative case or called unhealthy, where one is set for a positive case or called healthy. Similarly, for a range of age, if the patient is getting older, the health risk will be increasing. Therefore, the range of age from 18

TABLE 3.4: Experimented diseases

Code	Description
DIQ010	diabetes or sugar diabetes
KIQ022	weak or failing kidneys
MCQ160A	arthritis
MCQ160B	congestive heart failure
MCQ160C	coronary heart disease
MCQ160D	angina, also called angina pectoris
MCQ160E	heart attack
MCQ160G	emphysema
MCQ160O	COPD
MCQ160L	liver condition
HEQ030	hepatitis C
BPQ020	high blood pressure
BPQ080	high cholesterol

TABLE 3.5: Statistics of the dataset

Description	Number
Participants	9770
Attributes	2858
Diseases	30
Healthy case	5144
Unhealthy case	4626

to 65 is normalised by a new scale from 0 to 1.

Data Cleansing

Cleansing noisy data is a necessary task of mining data. In this study, the NHANES dataset also contains a large number of missing data. Therefore, the NHANES dataset needs to be corrected before applying the model. If all of the attributes have a small instance (the small number of participants), the study removed them from the dataset because they may have a negative effect on developing the model. Based on the statistics table from the real dataset, to ensure enough data for experiments as well as to reduce the maximum of the negative effect on retrieval information by missing data, all attributes that have more than 50% of missing data are removed. After cleaning the noisy data, the dataset consists of 318 attributes and is more beneficial for applying to the experiment. However, there is still a little missing data from 318 attributes so that we need another step to process this issue before applying the model. Because the dataset is normalized into the unique format by the binary presentation, the missing data was proceeded by replacing the average of all values in the respective attribute. After this processing, the dataset is available to apply for the experiments.

3.4.2 Performance Measurements

The standard metrics including accuracy, recall, precision and *F*-measure (Bowes et al. [2012]) were used to measure the experimental model's performance. These are defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

$$Percision = \frac{TP}{TP + FP} \tag{3.7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.8}$$

$$F - measure = \frac{2 * Recall * Percision}{Recall + Percision} \tag{3.9}$$

where TP is True Positive (*"Subject x is correctly labelled as belong to disease y"*), TN is True Negative (*"Subject x is correctly labelled as not belong to disease y"*), FN is False Negative (*"Subject x is incorrectly labelled as not belong to disease y"*), and FP is False Positive (*"Subject x is incorrectly labelled as belong to disease y"*).

Precision is the fraction of the number of samples correctly labelled as belonging to a disease among the total number of samples labelled as belonging to a disease. Recall is the fraction of the number of samples correctly labelled as belonging to a disease among the total number of samples that actually belong to a disease. Equal to one is the ideal result that corresponds to precision and recall. These two quantities can be combined into a single value by F-measure. Accuracy presents the correct classification rate of the model.

In this study, to ensure the reliability of evaluation, the k-fold ($k = 5$) validation approach was adopted. Five subsets were randomly generated from the NHANES dataset. Four of the five subsets would be used for training and the other one for testing in each experimental run. The final performance of the experimental model was counted based on the average performance of all five runs over the tests of twenty-two diseases. Both the

training and testing of the model had five rounds being used for the experiment. All five sub-sets of data were used to run the proposed model and then applied to the baseline model.

3.4.3 Baseline Model

The proposed model was evaluated by comparing it with the baseline model (Chen et al. [2016a]), representing state-of-the-art related research. The baseline model uses the semi-supervised learning algorithm called the semi-supervised heterogeneous graph on health (SHG-Health) to solve the classification problem with consideration of the relationship existing in the data's neighbourhood. Alternatively, our proposed model considers the semantic relations between data attributes and relations underlying the data, which was adopted to construct the heterogeneous information graph in experiments.

To build up the heterogeneous information graph, Chen et al. [2016a] used the general health examination (GHE) dataset covering data from 2005 to 2010 in Taipei City. The dataset was collected from 102,258 participants aged from 65. The baseline model has categorized the dataset with 230 attributes into three types of group data: physical test, mental assessment, and profile. A heterogeneous graph was set up with four different types of nodes: Record, Physical Test, Mental Assessment, and Profile. Record is an artificial node for representing individual records. This is an essential node of the baseline model because it is used for setting up the links of the graph. Each individual record presents for more than one node because each record of the GHE dataset contains data for different years from 2005 to 2010.

In contrast, the NHANES dataset used in this study covers data for only one year. This could point to an assessment of the performance result of the baseline model because the baseline model did not show whether or not the results are affected if the dataset has data for one year. In this study, the baseline model was rebuilt by using the NHANES dataset. Both our model and the baseline model were based on the same dataset to develop the binary classification model. Both models have the same goal of trying to make a significant result in predicting high-risk diseases. By applying a new proposal, this study tries to improve the performance of classification compared to the baseline model.

3.4.4 Results

Figure 3.3 presents the dataflow in an experimental design. First, the NHANES dataset was separated into two subsets: a training set and a testing set. The former was used to help develop the classification model, whereas the latter was used for evaluation of the proposed model. All variables of the dataset were presented as binary labels. A value of 1 indicates a positive case, and a value of 0 indicates a negative case. Based on the design model, all values were converted to binary. Then, a risk score value is computed to indicate a patient's health status. Finally, the results of the proposed model were compared to the SHG-Health model to show the percentage of improvement.

Thirteen diseases were tested in five rounds, as discussed. Tables 3.6, 3.7, 3.8, and 3.9 presented the experimental results in both the proposal model and SHG-Health model including precision, recall, and accuracy and F-measure. The percentage of precision in the proposed model showed a great

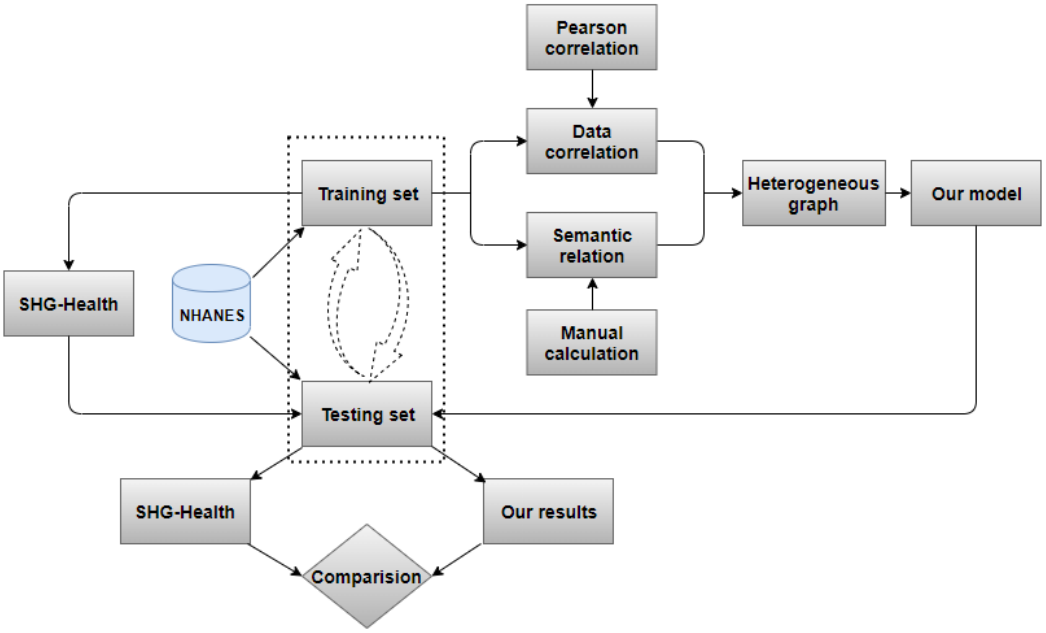


FIGURE 3.3: Experimental data-flow

improvement compared to the SHG-Health model. There were a few diseases that were not improving the performance of classification compared to the proposed model regarding recall and accuracy. Overall, although the percentage of recall and accuracy did not show improvement for some diseases in the proposed model, the proposed model outperformed the SHG-Health model in *F*-measure.

To justify the overall performance of the proposed model and SHG-Health model, the average of all diseases was present in Tables 3.10. The percentage change in performance was also used for clarifying the significant level of improvement achieved by the proposed model over the SHG-Health model. It is calculated by the following equation, where N is the number of diseases being observed in the experiment.

TABLE 3.6: Precision result of thirteen diseases, where the emphasised values indicate the superior performance in comparison.

Disease	Precision	
	Proposal model	**SHG-Health model**
KIQ022	**0.39918**	0.09882
MCQ160G	**0.32422**	0.12551
MCQ160D	**0.32913**	0.12248
DIQ010	**0.80073**	0.08815
HEQ030	**0.59007**	0.00987
MCQ160O	**0.66619**	0.13914
MCQ160B	**0.25806**	0.13696
MCQ160C	**0.46522**	0.11863
MCQ160E	**0.40931**	0.12981
MCQ160L	**0.58956**	0.09284
MCQ160A	**0.51509**	0.33932
BPQ020	**0.66636**	0.36759
BPQ080	**0.51177**	0.31788
Avg.	**0.50191**	0.16054

$$\%Chg = \frac{1}{N} * \sum_{i=1}^{N} \frac{result(proposal\ model) - result(baseline\ model)}{result(baseline\ model)} * 100\% \tag{3.10}$$

3.4.5 Discussions

The overall performance of the proposed model is better than that of the SHG-Health model. The result suggests that the proposed model has higher capability of processing sparse data compared with the baseline model. The training data set is sparse and non-balanced. The proposed model achieved promising results although dealing with such sparse data. However, the

TABLE 3.7: Recall result of thirteen diseases, where the emphasised values indicate the superior performance in comparison.

Disease	Recall	
	Proposal model	**SHG-Health model**
KIQ022	**0.43024**	0.23324
MCQ160G	**0.82700**	0.11167
MCQ160D	**0.53281**	0.14746
DIQ010	0.59232	**0.85820**
HEQ030	0.62748	**0.73551**
MCQ160O	**0.45330**	0.18103
MCQ160B	**0.63830**	0.23176
MCQ160C	**0.61373**	0.29357
MCQ160E	**0.61993**	0.24936
MCQ160L	0.25329	**0.34374**
MCQ160A	**0.73416**	0.66660
BPQ020	0.790084	**0.91995**
BPQ080	0.71754	**0.98800**
Avg.	**0.60232**	0.45847

SHG-Health model was recorded with relatively lower performance, especially, in Precision and F_1 measure results. The adoption of semantic and domain knowledge has made a significant impact on the success of the proposed model. The data was categorised into different groups based on the semantic and domain knowledge. The use of the Pearson correlation has also brought to the proposed model the ability of recognising the patterns underlying the data. With all such advantages, the prosed model was leveraged and overall it eventually significantly outperformed the baseline.

Although the result of the proposed model showed a significant improvement compared to the SHG-Health model overall, the performance of Recall and Accuracy for some diseases in the baseline model achieved better of classification. The SHG-Health model is designed to predict health risks

TABLE 3.8: Accuracy result of thirteen diseases, where the emphasised values indicate the superior performance in comparison.

Disease	Accuracy	
	Proposal model	**SHG-Health model**
KIQ022	0.78162	**0.90697**
MCQ160G	0.78415	**0.97144**
MCQ160D	**0.96374**	0.95542
DIQ010	**0.95709**	0.30568
HEQ030	**0.99251**	0.34476
MCQ160O	0.78666	**0.93204**
MCQ160B	0.75916	**0.93123**
MCQ160C	**0.95604**	0.88386
MCQ160E	**0.94924**	0.90406
MCQ160L	**0.96229**	0.83609
MCQ160A	**0.74359**	0.56926
BPQ020	**0.79554**	0.43991
BPQ080	**0.68894**	0.32439
Avg.	**0.85543**	0.71578

by mining a heterogeneous information graph generated from data. The heterogeneous information graph constitutes data with similar attributes and properties in a common neighbourhood. The model then tries discover health care / medical domain knowledge from the data constructed in the heterogeneous information graph form. Investigation revealed that the SHG-Health model tended to have obtained higher coverage of hypothesis space with a heterogeneous information graph constructed with an non-complex structure. As a result, in some diseases, the SHG-Health model performed better than the proposed model as shown in Table 3.7 and 3.8 when constructing the heterogeneous information graph using the NHANES dataset with data collected in only one year. This is also because of the sparseness of data. There are some disease in the dataset which have

TABLE 3.9: *F*-Measure result of thirteen diseases, where the emphasised values indicate the superior performance in comparison.

Disease	***F*-Measure**	
	Proposal model	**SHG-Health model**
KIQ022	**0.30101**	0.13624
MCQ160G	**0.43012**	0.11271
MCQ160D	**0.39693**	0.12820
DIQ010	**0.67798**	0.15984
HEQ030	**0.59204**	0.01947
MCQ160O	**0.37085**	0.14753
MCQ160B	**0.32307**	0.16690
MCQ160C	**0.52822**	0.16845
MCQ160E	**0.49208**	0.16804
MCQ160L	**0.35294**	0.14601
MCQ160A	**0.60085**	0.44798
BPQ020	**0.72234**	0.52498
BPQ080	**0.59288**	0.48094
Avg.	**0.49087**	0.21595

TABLE 3.10: Overall comparison of the proposal model to the baseline model, where the emphasised values indicate the superior performance in comparison.

	Precision	**Recall**	**Accuracy**	***F*-Measure**
The proposal model	**0.50191**	**0.60232**	**0.85543**	**0.49087**
SHG-Health model	0.16054	0.45847	0.71578	0.21595
Percentage Change	212.64%	31.38%	19.51%	127.31%

a small number of samples for the experiment. The non-balanced data led to different results for the experiment.

TABLE 3.11: Overall statistics for each threshold in the validation sets.

Threshold	**0**	**0.1**	**0.2**	**0.3**	**0.4**
% of links incl.	100%	24.36%	4.93%	1.34%	0.63%
Threshold	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
% of links incl.	0.21%	0.19%	0.34%	0.37%	0.97%

Furthermore, there are several parameters of importance related to the development of our model. Changing these coefficients may affect the results of our model. For that reason, the research needs to identify the best value to enhance the performance of our model. Firstly, it is related to the range values of the *Pearson correlation* coefficient. An analysis is needed to assess the best threshold value (γ) for the Pearson Correlation coefficient (ρ) because its threshold strongly affects the overall performance achieved by the model. If γ is set too large, the relevant information will not be sufficient for study. In contrast, if we set γ too small, more non-relevant information is included as noise. Therefore, we ran a program to test all range values of *Pearson correlation* from 0 to 1, scaling up by 0.1. The approach is to count how many links can be included in the information graph for each range values of the *Pearson correlation* coefficient. The results are presented in Table 3.11. The study found that $\gamma = 0.3$ gave the model the best and most stable performance. Table 3.12 illustrated the performance of the classification model by different the values of threshold γ. If the value of the threshold γ increased to 0.4 or even more, the performance of the classification model

could not be identified. As a result, γ = 0.3 was set for the *Pearson correlation*, which helped the construction of an information graph with a wealth of information to mine.

TABLE 3.12: Sample comparison of different the values of threshold γ in Experiments (BPQ080).

Threshold	Best averaged performance			
(γ)	Precision	Recall	Accuracy	F-measure
0.2	0.38628	0.9087	0.56367	0.52370
0.3	0.52231	0.71329	0.75472	0.59868
0.4	NaN	0	0.74139	NaN

Besides, the α, β, and θ coefficients play an essential role in deciding the final result of our model. In the experiment, an algorithm in Section 3.3.4 called the optimization procedure is applied in the training model to identify the best of α, β, and θ, where α + β =1, and θ is in the range from 0.1 to 0.9. Following this, all of these values are applied in the testing model to evaluate the results of the proposed model. Since the instance for each round is different, the performance of the result in the testing model for some diseases decreases compared to the training model. It is easy to understand that if there are not enough instances being applied to the model, the result will change. Because our model makes a comparison between the same and a different group, if the instances of the same group are not enough, the performance of our model is decreased. That is the reason why when we have the best value for the α, β, and θ coefficient in training, the results in the testing model may not be obtained as expected.

The coefficients of α and β in (*Eq*. 3.3) were designed to leverage the overall performance of the model by giving different considerations to the

latent health knowledge and domain health knowledge. Thus, the final determined values of α and β also reveal the importance of the latent health knowledge and domain health knowledge to the model. During the experiments, it was found out that when giving the latent health knowledge more consideration than the domain health knowledge (α holds a larger value than β), the proposed model would be powered with higher performance. Table 3.13 presents the experimental results on BPQ080 High cholesterol, one of the diseases studied in the experiments, with different values of α, β, and θ going through five rounds of K-fold. Three out of five rounds show that the latent health knowledge has presented a stronger influence than the domain health knowledge (in the second and fourth rounds they were tied).

TABLE 3.13: Sample comparison of latent health knowledge and domain health knowledge in Experiments (BPQ080).

	Threshold	Latent	Domain	Best performance			
Round	(θ)	(α)	(β)	Precision	Recall	Accuracy	F-measure
1	0.3	1	0	0.57385	0.56834	0.72230	0.57108
2	0.2	0.5	0.5	0.48743	0.73857	0.68707	0.58728
3	0.1	0.7	0.3	0.53445	0.75177	0.70705	0.62475
4	0.2	0.5	0.5	0.45833	0.75959	0.64598	0.57170
5	0.1	0.7	0.3	0.50477	0.76941	0.68231	0.60961

3.5 Summary

In recent years much effort has been invested in transforming healthcare services form traditional experience-base to evidence-base (Mirel and Carper [2014], Collins and Varmus [2015], Gardner et al. [2018]). Along the journey,

data mining and machine learning techniques have played an important role because the evidence in fact refers to the latent knowledge discovered from massive data collected from daily operations of healthcare services. Data mining, machine learning and knowledge engineering / management have provided technical foundation to the transformation of healthcare services, and more advanced techniques in these areas are in great demand, aiming at further improving the quality of healthcare services.

Answering the call, we constructed a health knowledge graph on this study using the National Health and Nutrition Examination Survey, a health examination dataset. Adopting the knowledge graph, a classification model was also introduced to predict the potential health risk for patients. The *Pearson correlation* coefficient was used to discover the correlation between data attributes. Health domain knowledge contained in the categorisation of disease was also adopted in the model to help build up the knowledge graph. Aiming at evaluating the proposed classification model, empirical experiments were performed, in which the proposed model was compared with a SHG-Health model implemented for a state-of-the-art model introduced by Chen et al. [2016a]. The experimental results showed that the proposed model outperformed the SHG-Health model, which was significant in Precision and F_1 measure.

Chapter 4

Knowledge Base for Medical Health Status Classification

In this Chapter, the study integrates the knowledge of the medical domain into the heterogeneous information graph, which is built in Chapter 3, to improve the accuracy of the patients' health risk prediction. The research mines the knowledge, which was extracted from titles, and abstracts of MEDLINE to discover how to assess the links between objects relating to medical concepts. A knowledge-base graph model then is developed for the prediction of a patient's health status. This knowledge-base graph has instances with influence rules, which helps the graph to generate relationships between nodes The results of the experiment shows that the knowledge-base model is superior to the baseline model and has demonstrated that the knowledge-base could help improve the performance of the classification model. The contribution of this study provides a framework for applying a knowledge-base in the classification model, which helps these models achieve the best performance of predictions. The experiment of this study affirmed that biomedical literature could assist in improving the performance of the classification model.

4.1 Introduction

The increasing amount of big data has opened many new challenges in data mining. Traditional data mining, which is performed at the data level, may not be highly effective in discovering knowledge for two reasons: firstly, each attribute at the data level has a unique label which has a limitation in discovering knowledge; secondly, it is challenging to infer implicit information among entities. In contrast, at the knowledge level, each attribute might have more than one label, which focuses on presenting information by semantic meaning rather than data. Therefore, mining at the knowledge level can help gather implicit details which can support the achievement of a higher level of knowledge discovery.

In terms of the knowledge base, Wang et al. [2017a] demonstrated that a medical knowledge base would have the capacity to improve the performance of discovering medical knowledge if it was integrated into the medical domain knowledge. Goh et al. [2016] argued that a knowledge base is useful in the clinical decision support system. In the medical area, MEDLINE[1] is a vital source because it contains a significant number of citations that are updated frequently in the medical field. However, most researchers focus on using MEDLINE to retrieve information. Xu and Wang [2013c] introduced a model to identify drug-disease associations by extracting documents from MEDLINE. Some researchers have suggested new methods of achieving a high quality in knowledge that has been discovered. Banuqitah et al. [2016] suggested a way to use multi-level learning from documents extracted in MEDLINE to improve the discovery of previously hidden useful knowledge. Therefore, MEDLINE would become more useful if it was

[1]https://www.nlm.nih.gov/bsd/medline.html

integrated into concepts of Medical Subject Headings (MeSH)[2] by instances which can be applied to applications of decision support.

In addition, applying the heterogeneous information graph (HIG) has significantly increased by the amount of research, thus providing an advantage for developing a classification model. Sun and Han [2013] demonstrated that mining the HIG could provide an effective way to improve the quality of mining data. Because of the advantage of the HIG, Ji et al. [2011b] introduced a classification algorithm from the HIG, based on ranking class algorithms. The results of the experiments showed that the proposed research is more accurate in generating classes as well as contributing a more meaningful ranking of objects in each class. In the medical domain, the use of the HIG, which was built from clinical data to predict the health risk status, has become a current significant topic. Perotte et al. [2015] constructed a HIG from electronic health record data to predict the progression of chronic kidney disease. The result showed that the performance of this proposed model is more accurate than other models. This advantage increased quickly with the volume of research (Xiong et al. [2018], Lei and Zhang [2019]) that used the HIG in predicting health risks.

It is clear that evidence-based medicine has a positive effect on improving the quality of healthcare. It aims to ensure that the clinician's opinion relies on available knowledge from the scientific literature. Cases et al. [2013] argue that medical knowledge plays a vital role in decision support for medical practitioners as well as in healthcare knowledge delivery. However, extracting and transforming the evidence-based medicine into the care processes may have significant challenges because of diverse information

[2]https://www.nlm.nih.gov/mesh/meshhome.html

contents and structures (Böckmann and Heiden [2013]).

4.2 Research Problem Formulation

The Chapter develops a knowledge graph to help improve the performance of the classification model. The approach of applying the graph base is used to solve the problem of classification. The graph distributes the information under the heading 'concepts' and 'relationships'. Each concept presents a node, and each relationship indicates a link among nodes. In this Chapter, the research tries to strengthen information for each node as well as to discover semantic knowledge within each node. Therefore, a knowledge base developed from the verification and evaluation of experts is embedded in the graph. The knowledge base helps to improve the reliability of the graph for discovering knowledge then the study applies this knowledge base to the development of a classification model.

To complete the task that develops a classification model based on the knowledge graph, the following definition is used:

Definition 4. 1. ***[Health Examination Records]***
A 3-tuple $\mathbb{R} = \{\mathcal{P}, \mathcal{A}, \mathcal{M}_{\mathcal{P}}^{\mathcal{A}}\}$ *is called the Health Examination Record, where* $\mathcal{P}$ *is a set of patients,* $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$, $|\mathcal{P}| = m$. $\mathcal{A}$ *is a set of attributes,* $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$, $|\mathcal{A}| = n$. *A matrix* $\mathcal{M}_{\mathcal{P}}^{\mathcal{A}}$ *is constructed by* $P \times A$. □

Definition 4. 2. ***[Heterogeneous Information Graph]***
A heterogeneous information graph is a 3-tuple, $G :=< V, E, M >$, *where* $V = \bigcup_i^t \mathcal{A}_i$, $\mathcal{A}_i = \{a_{i1}, a_{i2}, \ldots, a_{in_i}\}$, $i = 1, 2, ..., t$ *are t types of data objects and* $t \geq 2$, *E and M are the set of links between any two data objects V and the set of weight values by links, respectively.* □

The main target of this study predicts all nodes that have not been labelled in this graph. However, there is limited information for each node in this graph. Therefore, the research divides this graph into many subgraphs, so that each subgraph covers specific information which is defined as a class. In this case, each class presents a set of "Disease" type objects in the graph.

Definition 4. 3. ***[Disease Subgraph]***

Given a heterogeneous information graph $G :=< V, E, M >$ *and a disease* $k \in \mathcal{A}_d$, $\mathcal{A}_d$ *is the set of "Disease" type objects, a disease subgraph is a graph* $\acute{G} =< \acute{V}, \acute{E}, \acute{M} > \subseteq G$ *if* $\exists \mathcal{A}_d \subset V$. □

Then, we enrich information for each disease subgraph based on embedding instances. Instances help to improve and expand the semantic relationship between nodes which are learned from a specific knowledge based domain. Finally, we define our research problem in the following way:

Definition 4. 4. ***[Research Problem]***

Let $P = \{p_i, i = 1, \ldots, \imath\}$ *be a set of patients,* $\acute{G} = \{g(d_1), \ldots, g(d_K)\}$ *be a set of disease subgraph, where each d is a disease, K is the number of classes and* $P \in \mathcal{P}$. *Given a training set of patients* $P_t = \{p_j, j = \imath + 1, \ldots, m\}$, *the research problem is to learn a binary prediction function* $f(g^k|p)$ *and use it to classify* $p_i \in P$ *into* $\{g(d_k)\} \subset \acute{G}$ *for prediction of the patients' health status in terms of the set of diseases defined in* $\acute{G}$. □

4.3 Framework

4.3.1 Medical Subject Headings

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus which contains a set of terms naming descriptors as in Figure 4.1, also known as subject headings in a hierarchical structure. Each descriptor, has a short definition, connects to related descriptors and a list of similar terms. The 2018 version of MeSH includes a total of 28939 descriptors, 116909 descriptor terms and 244154 supplementary concept records (Tateisi [2019]). It was created by the National Library of Medicine in 1960 and covered aspects of medicine and healthcare. MeSH is used for indexing journal articles for MEDLINE. Each article in the MEDLINE database is assigned from six to fifteen subject headings from MeSH. It is updated annually to reflect current terminology usage. An advantage of MeSH is that it can help to search the most specific terms available in each article. Therefore, using MeSH to search articles in MEDLINE helps to obtain a high efficiency in searching information at various levels of specificity.

Definition 4. 5. ***[Medical Subject Headings]***
The Medical Subject Headings are $\mathbf{C} := < \mathrm{C}, \mathrm{R}_{\mathrm{C}}^{\mathrm{C}} >$, *where* $\mathrm{C} = \{c_1, c_2, \ldots, c_n\}$, n *is the number of concepts and* c_n *is a concept belong* C. $\mathrm{R}_{\mathrm{C}}^{\mathrm{C}}$ *is a matrix* $\mathrm{C} \times \mathrm{C}$. *The matrix is set up based on the semantic relationships defined by the medical subject heading.* □

4.3.2 MEDLINE

MEDLINE is a bibliographic database that has collected journal articles from academic journals in life sciences and biomedical information since 1966. It

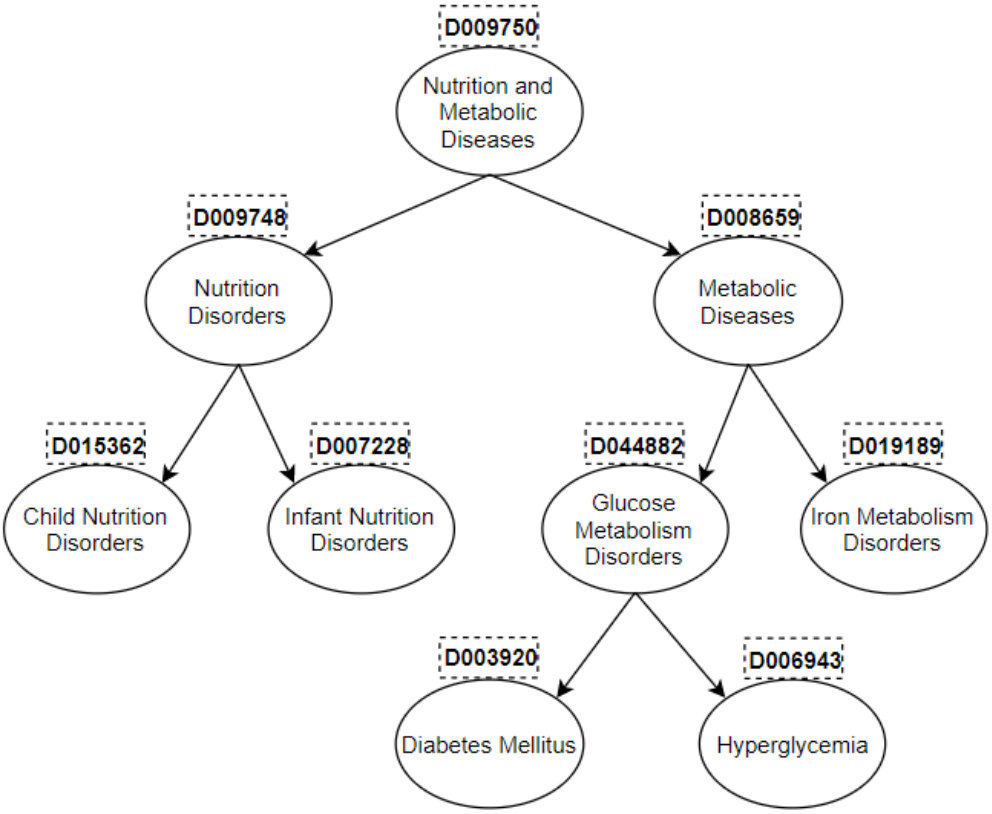


FIGURE 4.1: Medical Subject Heading

is produced by the National Library of Medicine in the United States. These academic journals cover medicine, veterinary medicine, nursing, dentistry, pharmacy, and health care. The database contains more than 27 million references which are selected from more than 5200 international publications in about 40 languages (Costa et al. [2018]). References are added to the database each week. MEDLINE uses Medical Subject Headings which give uniformity and consistency to the indexing of the biomedical literature for information retrieval. MEDLINE uses the PubMed[3] interface for free access on the Internet. Engines designed to search MEDLINE include author names, words in abstract and title of the article, date of publication, and MeSH terms. Figure 4.2 presents an example of MEDLINE. As shown in Figure, each citation has an identity that links a few specifical descriptors, terms or concepts (e.g. the descriptor "Diabetes Mellitus, Type 2") from MeSH. It is also associated with one or more titles and abstracts of articles related to diabetes mellitus. Journal articles are selected based on the recommendations of the Literature Selection Technical Review Committee from advisory committees of both external and internal experts. Therefore, the function of MEDLINE is to be an important resource for biomedical researchers around the world.

Definition 4. 6. ***[MEDLINE]***

MEDLINE is the set of document $\mathfrak{D} = \{\mathfrak{d}_1, \mathfrak{d}_2, \ldots, \mathfrak{d}_m\}$*, where m is the number of documents in* $\mathfrak{D}$*.* $\mathfrak{d} := \langle \mathbf{T}, map(\mathfrak{d}) \rangle$*, where* $\mathbf{T} = \{t_1, t_2, \ldots, t_z\}$ *is a set of terms from d, and* $map(\mathfrak{d}) \longrightarrow \mathbf{C}_\mathfrak{d} \subset \mathbf{C}$*.* □

[3]https://www.ncbi.nlm.nih.gov/pubmed/

```
<MedlineCitation Owner="NLM" Status="MEDLINE">
<PMID Version="1">23575045</PMID>
<DateCreated>
<DateCompleted>
<Article PubModel="Print">
<Journal>
<ArticleTitle>Abdominal circumference as a screening measure for Type 2
<Pagination>
<Abstract>
<AbstractText Label="BACKGROUND" NlmCategory="BACKGROUND">No comparativ
<AbstractText Label="OBJECTIVE" NlmCategory="OBJECTIVE">This study aims
<AbstractText Label="METHOD" NlmCategory="METHODS">On 187 adult males a
<AbstractText Label="RESULTS" NlmCategory="RESULTS">The prevalence of T
<AbstractText Label="CONCLUSION" NlmCategory="CONCLUSIONS">Waist circum
</Abstract>
<AuthorList CompleteYN="Y">
<Language>eng</Language>
<PublicationTypeList>
</Article>
<MedlineJournalInfo>
<ChemicalList>
<CitationSubset>IM</CitationSubset>
<MeshHeadingList>
<MeshHeading>
<DescriptorName MajorTopicYN="N" UI="D003924">Diabetes Mellitus, Type 2
<QualifierName MajorTopicYN="Y" UI="Q000175">diagnosis</QualifierName>
</MeshHeading>
<MeshHeading>
<MeshHeading>
</MeshHeadingList>
</MedlineCitation>
```

FIGURE 4.2: MEDLINE

4.3.3 Population Medical Knowledge Graph

In this study, medical subject headings are considered as subgraphs to populate knowledge from MEDLINE through instances. Figure 4.3 shows an example of an association between MeSH and MEDLINE. Concepts and terms related to a kind of disease (for example, type 2 diabetes mellitus) from MeSH are linked with a number of citations from MEDLINE that are assigned to type 2 diabetes mellitus. These connections with the concepts and terms associated to type 2 diabetes mellitus are grouped together to build up a subgraph. Each subgraph corresponds to a subject or a type of disease. Figure 4.4 presents an example of three subgraphs regarding liver

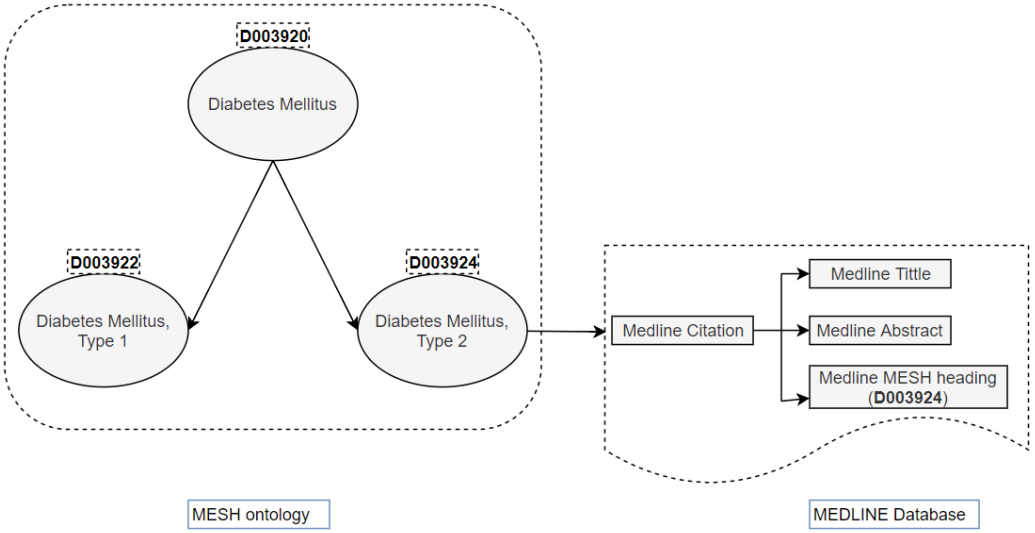


FIGURE 4.3: Association between MESH and MEDLINE

cancer, kidney cancer and diabetes. These subgraphs are built by discovering the knowledge from MEDLINE and are an innovative approach for gaining the accuracy of a classification model. To populate the knowledge base for these subgraphs, a mapping function is needed to map the clinical data and MEDLINE.

As indicated by the discussion above, to map the observation data and MEDLINE, the study maps between MeSH and International Classification of Diseases, Tenth Revision (ICD-10)[4].

Definition 4. 7. ***[ICD-10]***

The ICD-10 is the set of disease $\mathbf{D} = \{d_1, d_2, \ldots, d_m\}$*, where* m *is the number of diseases and* d_m *is a disease.* □

Given a set of diseases **D** and a set of concepts **C** from Definition 4.5. For each $d \in \mathbf{D}, Map(d \longmapsto \mathbf{C}) = \{c\} \subset \mathbf{C}$. For each $c \in \mathbf{C}, Map(c \longmapsto \mathbf{D}) =$

[4]https://www.cdc.gov/nchs/icd/icd10.htm

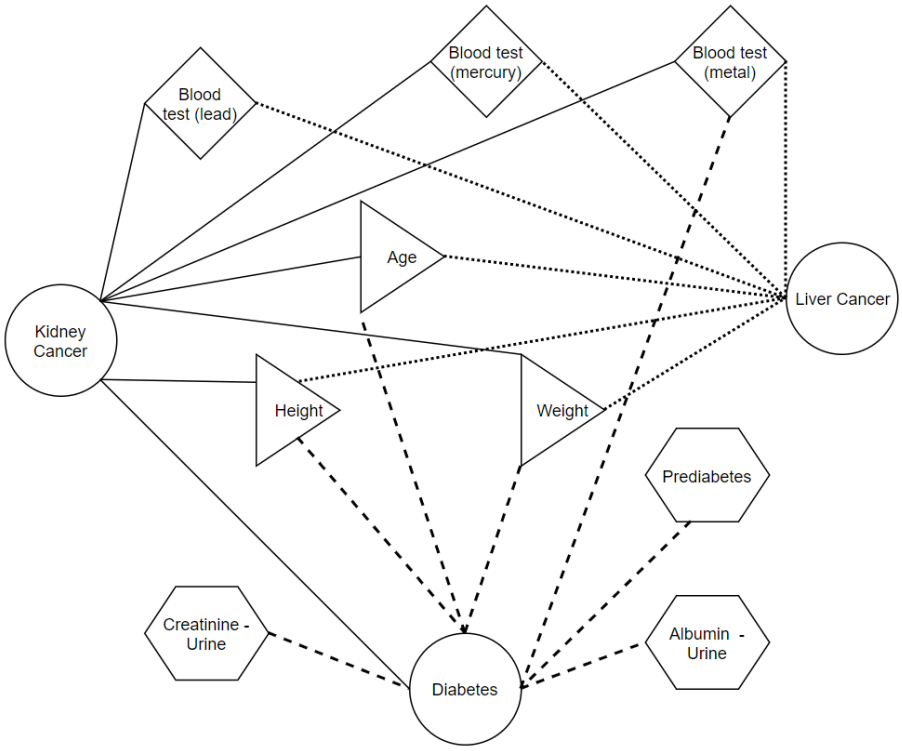


FIGURE 4.4: An example of subgraphs: there are three kinds of disease corresponding to three subgraphs. Each subgraph has a different number of nodes and type of links. The various subgraphs may have the same node together. Each node may belong to a different object.

$\{d\} \subset \mathbf{D}$. Given c_i and d_j, a link is generated based on the following:

$$Link(c_i, d_j) = \begin{cases} 1, & \text{if } |Map(d \longmapsto \mathbf{C})| > 0 \vee |Map(c \longmapsto \mathbf{D})| > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

The Unified Medical Language System (UMLS)[5] Metathesaurus is used as a standard to identify relationships among these data. UMLS consists of various terms of different sources in biomedicine and health care which were created by the U.S. National Library of Medicine (NLM) in 1986. If a

[5]https://www.nlm.nih.gov/research/umls/index.html

descriptor's coding (e.g. D003920 identified as a Diabetes Mellitus, which is showed in Figure 4.3) in MeSH and a disease coding (e.g. E11 is defined for Diabetes Mellitus) in ICD-10 have the same concept in UMLS, a link is established to connect MeSH and ICD-10. The mapping between disease names and code is manually classified by the U.S. National Library of Medicine (NLM). To identify links though mapping MeSH and ICD10, an **Algorithm** 4.1 is introduced as follows:

Algorithm 4.1 Algorithm for identifying links between MeSH and ICD10

1: **INPUT**: MeSH and ICD-10
2: **OUTPUT**: The set of links between MeSH and ICD-10
3: **Initialise** $\mathrm{L} = \varnothing$;
4: **for all** $c_i \in$ **C do**
5: **for all** $d_j \in$ **D do**
6: **if** $Link(c_i, d_j) = 1$ **then**
7: $\mathrm{L} = \mathrm{L} \cup \{\langle c_i, d_j \rangle\}$;
8: **end if**
9: **end for**
10: **end for**
11: **return** L

Algorithm 4.1 first uses a loop to check all diseases belonging to MeSH. Then, another loop is used to check all diseases in ICD10. If the same disease exists between MeSH and ICD10, a link will be set up and stored in set L. Then, repeating all values in the two sets MeSH and ICD10 will identify the rest of the links. Finally, L is returned with a set of links between MeSH and ICD10. The algorithm uses two nested loops for identifying the links between MeSH and ICD10. The complexity of the algorithm is $O(n^2)$.

Based on the mapping, a triangle of NHANE, MeSH and MEDLINE creates a knowledge graph in the medical domain. Each specific disease coding corresponds to a subgraph which is populated with knowledge from a large

number of articles from MEDLINE. Titles and abstracts of the articles were selected from published research that has been assessed by experts. Mining the extracted information to populate knowledge for this subgraph has potential value. Based on the subgraph, a specific knowledge domain can be obtained. The acquired knowledge may be used as medical evidence which is embedded in the heterogeneous information graph to estimate the performance of the classification model.

Definition 4. 8. ***[Knowledge Graph Base]***

The Knowledge Graph Base is a 3-tuple $\mathbb{KG} := \langle \mathcal{C}, \mathcal{R}, \mathbf{G}_{\mathcal{C}}^{\mathcal{R}} \rangle$*, where*

- $\mathcal{C} := \langle \mathbf{C}, \mathcal{I} \rangle$, $\epsilon := \langle c, \mathcal{I}_c \rangle$, *where* $\mathcal{I}_c \subset \mathcal{I}, c \in \mathbf{C}$. $\mathcal{I}$ *is the universal set of instances.*
- $\mathcal{R} = \{r_1, r_2, \ldots, r_q\}$ *is the set of all relation types in a knowledge graph, and* q *is the number of relations.*
- $\mathbf{G}_{\mathcal{C}}^{\mathcal{R}}$ *is a graph that is generated by* $\mathcal{R}$ *and* $\mathcal{C}$

□

Based on the definition of knowledge Base, the study needs to perform an important task to learn instances from $\mathfrak{D}$ based on $\mathbf{T}$ and *map(*$\mathfrak{d}$*)*. These instances which are associated with concepts $\mathbf{C}_{\mathfrak{d}} \subset \mathbf{C}$, play an essential role in building a knowledge graph $\mathbb{KG}$.

Assume that both $\mathbf{C}$ and $\mathfrak{D}$ have k different subjects. Each k subject has n concepts. MeSH is presented as $\mathbf{C}_i = \{c_{i1}, c_{i2}, \ldots, c_{in}\}, i = 1, 2, ..., k$. Where MEDLINE is indicated as $\mathfrak{D}_i = \{\mathfrak{d}_{i1}, \mathfrak{d}_{i2}, \ldots, \mathfrak{d}_{im}\}, i = 1, 2, ..., k$ and m is the number of documents. To learn instances from any subject k for populating knowledge of subgraph corresponding subject k, $\mathcal{C}_k := \langle \mathbf{C}_k, \mathcal{I}_k \rangle$, where $\mathcal{C}_k \subset \mathcal{C}$. $\mathcal{I}_k$ is learned by mapping between $\mathbf{C}_k$ and $\mathfrak{D}_k$.

$$f(\mathcal{I}_k) = \sum_{j=1}^{z} (c_{(k)} \longmapsto t_{(k)j}) \tag{4.2}$$

These instances $f(\mathcal{I}_k)$ are integrated with edges in the heterogeneous information graph for improving the accuracy of the classification model.

4.3.4 Applying Knowledge Base to the Classification Model

In the work presented in Pham et al. [2018], a heterogeneous information graph was built to deal with the problem of health risk prediction. Based on the work, the study tries to achieve the optimum knowledge discovery by applying the strongest effective factors in building the heterogeneous information graph. A heterogeneous information graph is a 3-tuple, $G :=< V, E, M >$ with an object mapping function $\varphi : V \rightarrow A$ and a link type mapping function $\psi : E \rightarrow R$. A heterogeneous graph consisting of different types of nodes was built based on health examination records. *Pearson correlation* was used to identify the connection between two nodes in the graph. By using h attributes to calculate the coefficient values of the *Pearson correlation*, the strength of the relationship between two nodes in the heterogeneous graph was effectively demonstrated, denoting a *Pearson correlation* coefficient value by ρ, representing the relationship between two vertices. The connection is valid if $\rho(v_1, v_2) \geq \gamma$, where γ is a threshold defining the validity of object connection. The range of the ρ is from -1 to 1; if the coefficient value approaches 1, two vertices v_1 and v_2 are strongly correlated. However, if the coefficient value is in the range of nearly -1, v_1 and v_2, they are weakly correlated. Overall, the larger the value is, the stronger is the relationship between v_1 and v_2. Moreover, in the study, *semantic relations* in

Section 3.3.2 from Chapter 3 were applied to advance the accuracy of information retrieval. To perform risk prediction for each patient based on the heterogeneous graph, each type of information node was regarded as a semantic class. The study, therefore, built a model to predict the risk of each disease belonging the various classes by comparing the effect among these classes.

As the discussion above suggests, a heterogeneous information graph was constructed based on the *Pearson correlation* and the *semantic similarity*. Later, the weight of edges $\rho(v_1, v_2)$ in the heterogeneous information graph was normalised by using instances $f(\mathcal{I}_k)$ (Eq.(4.2)). The weight of edges was a critical characteristic of a network that was considered in the complex system by Supriya et al. [2016] for the detection of the epilepsy syndrome. After populating the knowledge for the heterogeneous information graph through instances, a new knowledge graph was generated that was called the Knowledge Based Heterogeneous Information Graph (KB-HIG). Then, a classification model was developed based on the KB-HIG for predicting health risk status. The formula of normalisation is:

$$f(V_k) = \sum_{i=1}^{n} \rho(v(k), v_i) * f(\mathcal{I}_k) \tag{4.3}$$

Assume $\{e_{i1}, e_{i2}, \ldots, e_{iq}\} \subseteq E$, $i = 1, 2, ..., k$, was the set of links between any two data objects of a knowledge graph, where k is the type of subjects and q is the number of the leaning instances. Eq.(4.3) was presented:

$$f(V_k) = \sum_{i=1}^{q} e(k)_i \tag{4.4}$$

Basically, this function (Eq. 4.4) helps to populate the knowledge for a

graph through instances which learned from MEDLINE. All instances $\mathcal{I}_k$ were identified through a matrix between NHANES and MEDLINE. First, the study considered all of the attributes from NHANES being independent terms regarding a type of subject. Secondly, by extracting titles and abstracts from MEDLINE corresponding to a kind of disease that was the same type of subject with NHANES, a list of terms concerning a kind of disease was generated through the information network as a subgraph. A matrix was created by combining terms between NHANES and MEDLINE. Figure 4.5 illustrates an example of the connection between NHANES and MEDLINE for diabetes. If any variable of NHANES contains any term from the diabetes information network in MEDLINE, these terms were marked this variable. All mapping between terms of NHANES and terms of MEDLINE were called learned instances.

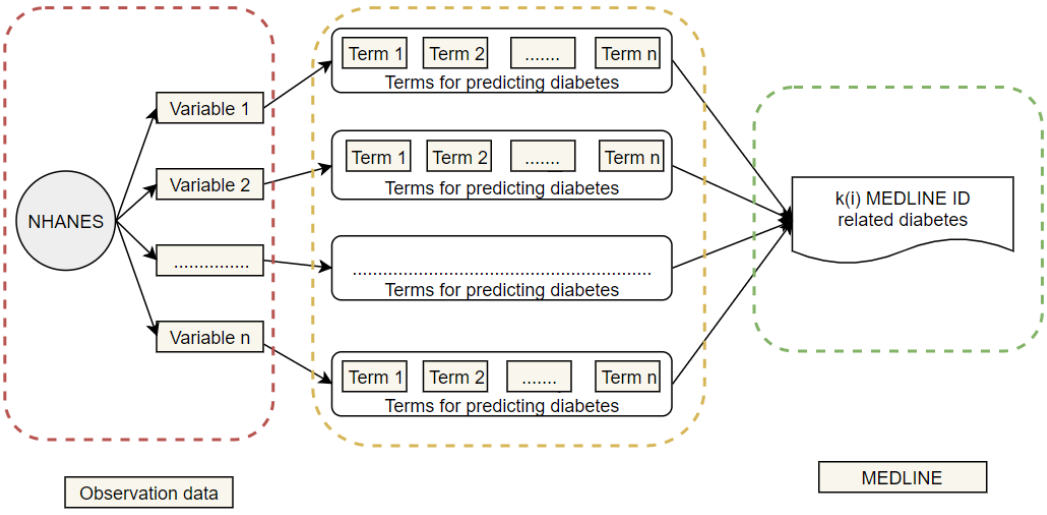


FIGURE 4.5: Mapping variable from NHANES to terms from MEDLINE

In this study, the study applied the word2vec algorithm suggested by Mikolov et al. [2013] to identify the value for each instance. Word2vec can transform a large corpus of text into a vector space with several hundred

dimensions. Every single word from the corpus is assigned a corresponding vector. By utilising the word2vec algorithm, Zheng and Callan [2015a] and Zheng and Callan [2015b] succeeded in extracting and calculating the weight among terms. This technique is also called embedding and can measure semantic similarity among terms as well as recognise similar neighbours for a given term. Obtaining the weight of among terms has a significant effect on the task of mining knowledge because the weight of among terms could help promote semantic similarity among terms. These benefits might contribute to improving the accuracy of the classification model. The ranges value for each term is from zero to one. For example, if a term does not exist or is not associated with the given term (e.g. diabetes), the value is nearly zero. In contrast, the value is approximately one if the term is more related to diabetes.

By using the availability of the KB-HIG, the research built a function that can formalise the profile of a healthy (unhealthy) patient concerning a disease object x:

$$f(x) = \sum_{i=1}^{l} \sum_{h=1}^{q} v_i \times e_h \times \alpha + \sum_{j=1}^{l} \sum_{h=1}^{q} v_j \times e_h \times \beta \tag{4.5}$$

where $\varphi(x) = \varphi(v_i)$ and $\varphi(x) \neq \varphi(v_j)$. Each element v_i or v_j is an attribute $a \in \mathcal{A}$. e_h is the weight of links, which is normalized through a knowledge graph. α and β are two coefficients adopted to clarify the contribution of *latent health knowledge* and *domain health knowledge* in the classification model.

4.4 Experimental Result and Analysis

4.4.1 Experimental Settings

In this research, diabetes mellitus was selected for evaluating the proposed model throughout two data sets to ensure the consistency of the performance. Diabetes mellitus was one of the most severe health problems and has caused 79% of deaths for people under the age of 60 by the statistical report of the World Health Organization (Shakeel et al. [2018]). Many researchers considered using this disease over the past decade for evaluating the performance of predictive models. Luo's experiment (Luo [2016]) investigated this disease in generating rules for explaining the results of the predictive models. Diabetes mellitus was also considered for estimating the performance of experiments by Boytcheva et al. [2017]. They extracted entities from a big collection of outpatient records using frequent patterns mining.

Figure 4.6 presents the dataflow in an experimental design. A k-fold ($k = 5$) validation approach was adopted for the experiment. Four of the five subset data extracted from the knowledge base were used to train our model. Later, one of the five subset data was used to evaluate our model. With evaluation based on ground truth, this project uses standard metrics accuracy, recall, precision, and F1 measure to evaluate the performance of the model (Bowes et al. [2012]). The result was compared to Chen's work (Chen et al. [2016a]) called a Semi-supervised Heterogeneous Graph-based Algorithm for Health (SHG-Health) model. Moreover, to increase the reliability and evidence, the study integrated the knowledge base with the Heterogeneous information Graph (HIG) model (Pham et al. [2018]) and also

with the SHG-Health model to compare the effect of the knowledge base. Integrating the knowledge base into two original classification models was used to answer these two questions:

Q1) What is the impact of the knowledge base on the performance of the classification model?

Q2) Does the knowledge base make a contribution to classification models?

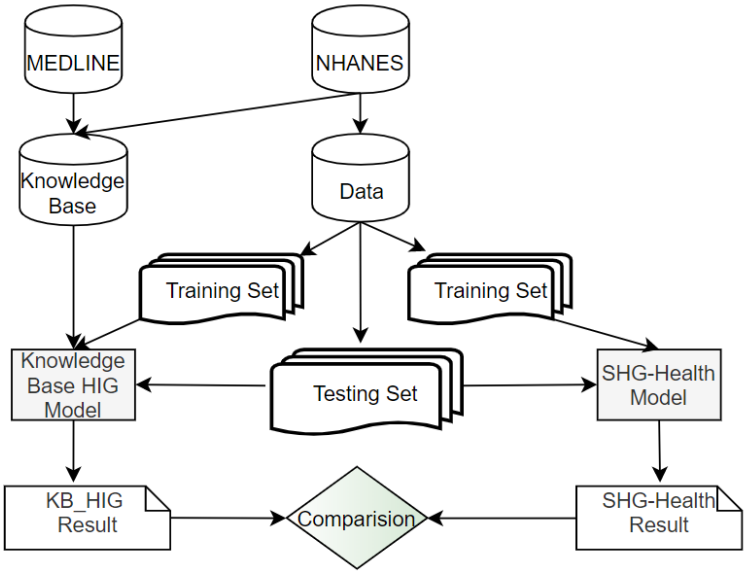


FIGURE 4.6: Experimental data-flow

One of the most important tasks of this experimental design is to map the observation data and medical knowledge. The study processed both MEDLINE and MeSH from XML format into a standard structure data before applying this data to train the model. The new construction helped to provide efficient access to the concepts and relationships in MeSH as well as to each document in MEDLINE. To solve the issues, a XML parser was written by using Java programming language to create a new structure of the database in MySQL under a table. This step made it convenient to extract information from MEDLINE by using MeSH for information retrieval.

After mapping MeSH and ICD-10, 153 types of diseases were detected associated with 3257 patients in NHANES. Table 4.1 shows the statistics of the top 10 mappings between ICD and MeSH. The research needs to extract titles and abstracts of MEDLINE ID correlated to diabetes mellitus to generate instances in normalising the weight of edges in the graph. After mapping ICD-10 and MeSH, the descriptors' coding of MeSH (D003924) was identified in relation to diabetes mellitus, which is shown in Table 4.1. The descriptors' coding was used to extract papers linked to diabetes mellitus in the MEDLINE database. 99785 papers associated with diabetes mellitus were used for the experiment.

TABLE 4.1: Top 10 mapping disease between ICD-10 and MeSH

ICD-10 Code	**Description code**	**Patient**	**Percent**	**MeSH code**
I10	Essential hypertension	2421	17.75%	D006973
E11	Diabetes mellitus	924	6.77%	D003924
J45	Asthma	544	3.99%	D001249
F32.9	Major depressive disorder	488	3.57%	D003863
K21	Gastro-esophageal reflux	445	3.26%	D005764
F41.9	Anxiety disorder	389	2.85%	D001008
E03.9	Hypothyroidism	340	2.49%	D009230
K30	Functional dyspepsia	175	1.28%	D004415
T78.40	Allergy	159	1.66%	D006967
I50.9	Heart failure	122	0.89%	D006333

The extracted data was used to generate a list of terms related to diabetes mellitus. The weight of each term was calculated through the word vector space model. In this study, word2vec algorithm (Mikolov et al. [2013]) was used to convert the extracted data into vector space. This algorithm helped to calculate the semantic relationship between all of the terms associated

with diabetes mellitus. Before using this extracted data for word vector space model, this study removed all stop-word and steam-word to improve accuracy information retrieval as well as ensure enough information in calculating the weight for each term.

After running the word vector space model, each term of diabetes mellitus is assigned a value. Table 4.2 presents the weight of the top 10 terms. Finally, a matrix is created by linking these term to variables in NHANES. The cooperation terms are generated based on this matrix. The value of each instance was identified by an association between a variable in NHANES and a list of terms for diabetes mellitus. For example, if a list of terms for diabetes mellitus does not exist in a variable concerning the attribute of age, the instance value of the age variable is set to 0. In contrast, the weight of an instance for a variable is equal to the weight of each term that link to this variable. Table 4.3 presents the top 10 variables of NHANES that have the highest weight.

TABLE 4.2: Top 10 terms for diabetes mellitus after mining

Term	**Weight**
type	0.80688
mellitus	0.80014
non	0.76048
dependent	0.75367
patient	0.75254
study	0.72770
niddm	0.70903
control	0.68019
insulin	65438
subject	0.63372

In this study, the proposed model called the Knowledge Base Heterogeneous Information Graph (KB-HIG) model is compared with three models

TABLE 4.3: Top 10 attributes from NHANES have a strong effect to diabetes mellitus

Variables	Weight
Taking insulin	0.65438
Age	0.56882
Glucose refrigerated serum	0.47249
Blood test	0.44453
Work activity	0.35117
Difficulty concentrating	0.34173
Cotinine Serum	0.33450
Mean cell volume	0.30563
Number of adults in household	0.26273
Total protein	0.24524

to evaluate the effect of the result if the researcher considers using a knowledge base in diagnosing the specific problem. These models include the HIG model, the SHG-Health model, and the Knowledge Base Semi-supervised Heterogeneous Graph-based Algorithm for Health (KB-SHG-Health). The HIG model used both *semantic similarity* and *Pearson correlation* to build a classification model. In contrast, the SHG-Health Model used the semi-supervised learning algorithm that conceded only the neighbourhood node in the heterogeneous graph to deal with the classification. By adopting the knowledge base in these two models, the study hopes to achieve improved results compared to the original model.

4.4.2 Dataset

The study used the observation data of the National Health and Nutrition Examination Survey (NHANES)[6] and the National Ambulatory Medical

[6]https://www.cdc.gov/nchs/nhanes/index.htm

Care Surveys (NAMCS)[7] for training sets and testing sets to evaluate the proposed model. The NHANES dataset has hundreds of available parameters which collect a wide range of health assessments such as lab tests, physical examinations, and personal habits. NHANES dataset contains 9770 participants with more than 2585 attributes. To ensure enough data for the experiment, only 318 attributes are used in the training model as well as testing results because of the missing data. The NAMCS dataset has a number of variables regarding the patient's smoking habits, the physician's diagnosis, the diagnosis, and prescription status as well as the demographic information on patients (for example. age, sex, weight, height, and race). The NAMCS dataset contains 32281 patients with 440 attributes. 164 attributes are used in the experiment after removing the missing data.

Before applying these data for training models, data preprocessing, including data cleaning and data normalisation were conducted because the dataset has a different type of data and contains a large amount of missing data. All variables of the dataset were presented as a binary label. A value of "1" indicates a positive case and a value of "0" a negative case. Based on the designed model, all data values were normalised as lying between 0 and 1 for validity in the experiment. For example, a nominal data type such as gender was converted into zero and one from male and female and an ordinal data type such as general health condition was converted into 0, 0.5 and 1 comprising poor, good and excellent. With the data of that range, results (50 to 150) such as blood test values were converted into the format of minimum and maximum [0, 1]. The step also provided a standard to identify positive and negative cases where 0 was set for negative situations

[7] https://doi.org/10.3886/ICPSR31482.v3

(unhealthy), and 1 was set for positive cases (healthy). The missing data were replaced by the average of all values in the respective attribute.

4.4.3 Baseline Model

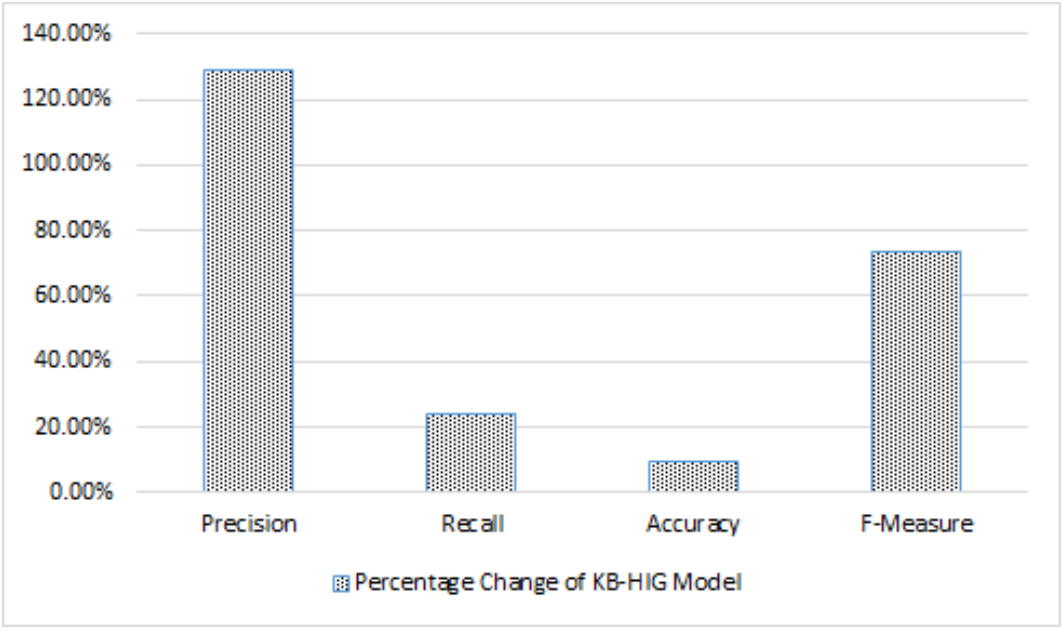
In this study, the research uses both the HIG model being developed in Chapter 3 and Chen's Model (Chen et al. [2016a]) to evaluate the performance of the proposed model. In Chapter 3, the study suggests a heterogeneous graph classification model (HIG model) by using both *semantic similarity* and *Pearson correlation* to build a classification model. In contrast, Chen's Model (SHG-Health Model) used the semi-supervised learning algorithm that considered only the neighbourhood node in the heterogeneous graph to deal with the classification. The result has demonstrated that the achieved performance of the HIG model overcome the SHG-Health Model. By adopting knowledge base in these two models, the study hopes to achieve more improvement compared to the original model.

4.4.4 Experimental Results

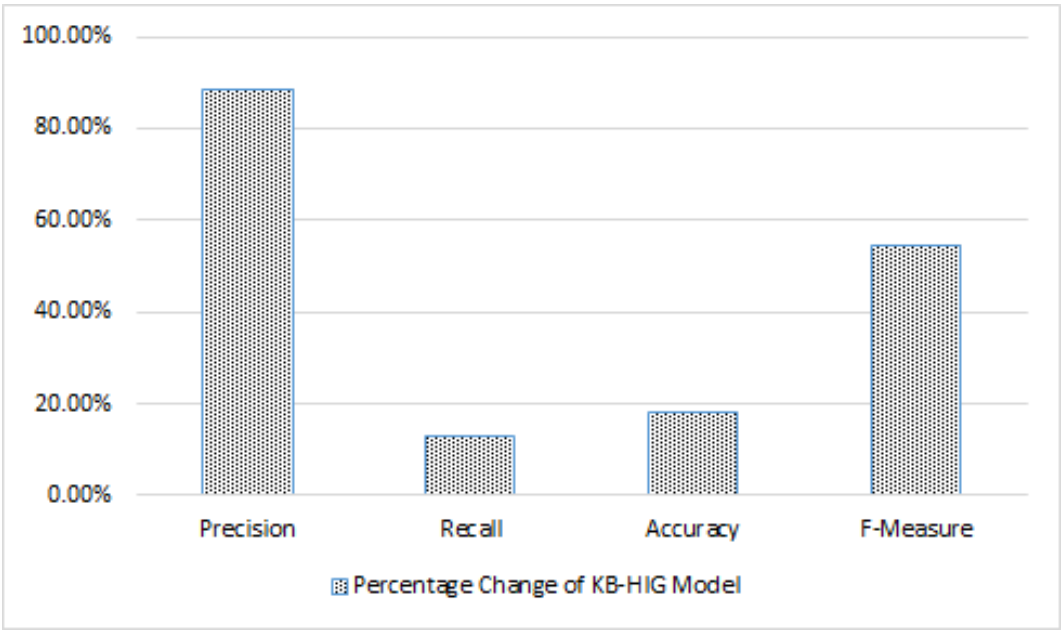
TABLE 4.4: Comparison between KB-HIG Model and SHG-Health model by NHANES, where the emphasised values indicate the superior performance in comparison.

	Precision	**Recall**	**Accuracy**	***F*-Measure**
KB-HIG model	**0.65982**	**0.68121**	**0.94243**	**0.65541**
SHG-Health model	0.28781	0.54998	0.86051	0.37753
Percentage Change	129.25%	23.86%	9.52%	73.60%

To apply a knowledge base in a classification model to advance the performance of prediction, the study completed a task to map the observation



(A) NHANES



(B) NAMCS

FIGURE 4.7: The improvement of the KB-HIG Model compared to SHG-Health model

TABLE 4.5: Comparison between KB-HIG Model and SHG-Health model by NAMCS, where the emphasised values indicate the superior performance in comparison.

	Precision	**Recall**	**Accuracy**	***F*-Measure**
KB-HIG model	**0.53060**	**0.64206**	**0.87097**	**0.58050**
SHG-Health model	0.28125	0.56780	0.73726	0.37563
Percentage Change	88.65%	13.08%	18.14%	54.54%

data in the NHANES dataset and MEDLINE based on MeSH and ICD-10. The statistic of mapping is presented in Section 4.4.1. After the mapping, the research performed a task of language processing to extract documents linked to diabetes mellitus. The obtained data were used to integrate into the classification model for predicting diabetes mellitus. The result of the experiment is indicated in Table 4.4, Table 4.5, and Figure 4.7. The precision of the KB-HIG model was a significant improvement compared to the baseline model in both two datasets. The result of recall and accuracy has improved by approximately 20 per cent. The percentage change in performance was identified based on the Equation 3.10. Overall, the experimental performance of KB-HIG model is better than SHG-Health model, which is justified through the vale of F-Measure by the NHANES and NAMCS dataset with nearly 75 per cent and around 55 per cent, respectively.

4.4.5 Discussions

The results obtained from the study demonstrated that the accuracy of the predicted outcome is improved by applying the knowledge base in the classification model. Using the knowledge base in the classification model helped to generate the semantic relationship among other objects. In this study, the research converted the extracted data regarding diabetes mellitus to the

word vector, which could identify the effect among other keywords or terms based on the vector space model. The technique finds out terms that are similar as well as being different to others by calculating the distance between the two terms. The attributes do not correlate to the topic (as liver cancer) being removed based on the threshold value that was calculated through the word2Vec algorithm. This approach helped the model obtain high performance in predicting disease.

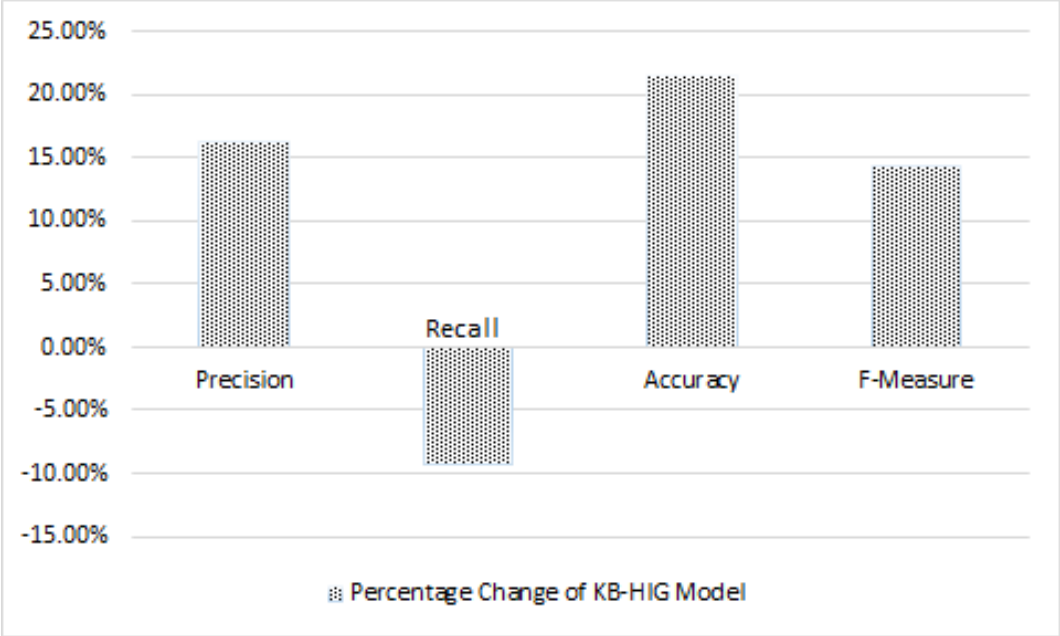
TABLE 4.6: Comparison between KB-HIG Model and HIG model by NHANES, where the emphasised values indicate the superior performance in comparison.

	Precision	**Recall**	**Accuracy**	***F*-Measure**
KB-HIG model	**0.65982**	0.68121	**0.94243**	**0.65541**
HIG model	0.56742	**0.75101**	0.77476	0.57282
Percentage Change	16.28%	-9.25%	21.64%	14.42%

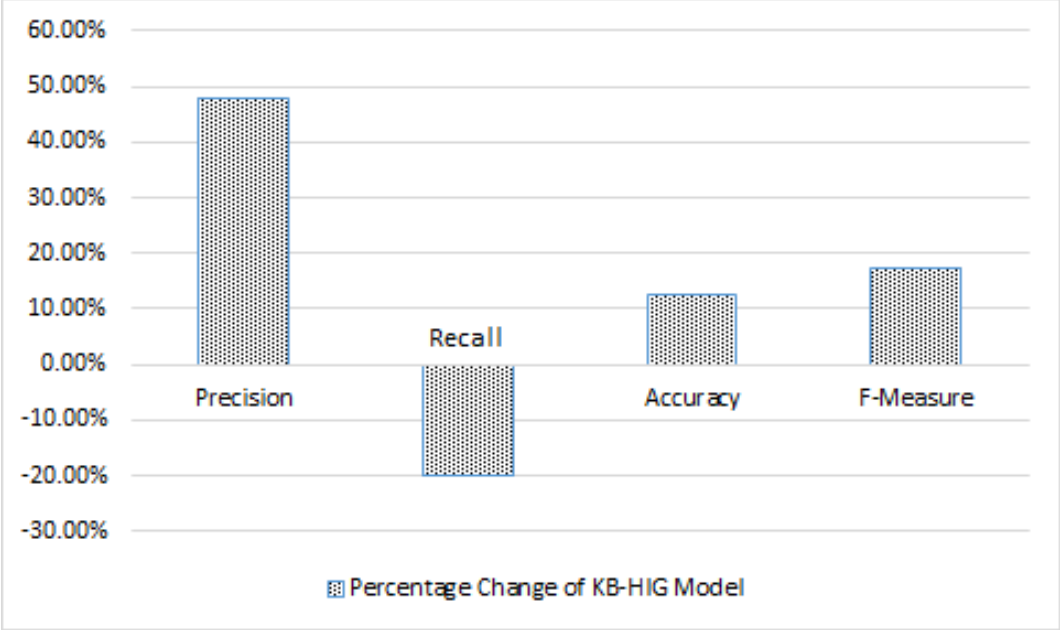
TABLE 4.7: Comparison between KB-HIG Model and HIG model by NAMCS, where the emphasised values indicate the superior performance in comparison.

	Precision	**Recall**	**Accuracy**	***F*-Measure**
KB-HIG model	**0.53060**	0.64206	**0.87097**	**0.58050**
HIG model	0.35870	**0.80332**	0.77247	0.49552
Percentage Change	47.92%	-20.07%	12.75%	17.15%

To increase the reliability as well as to provide more evidence about the advantage of this study, we expanded our experiment by applying the knowledge-base in other classification models. We aim to improve our claim that the knowledge-base could have a significant impact on different classification modes. There are two questions that we aim to answer.



(A) NHANES



(B) NAMCS

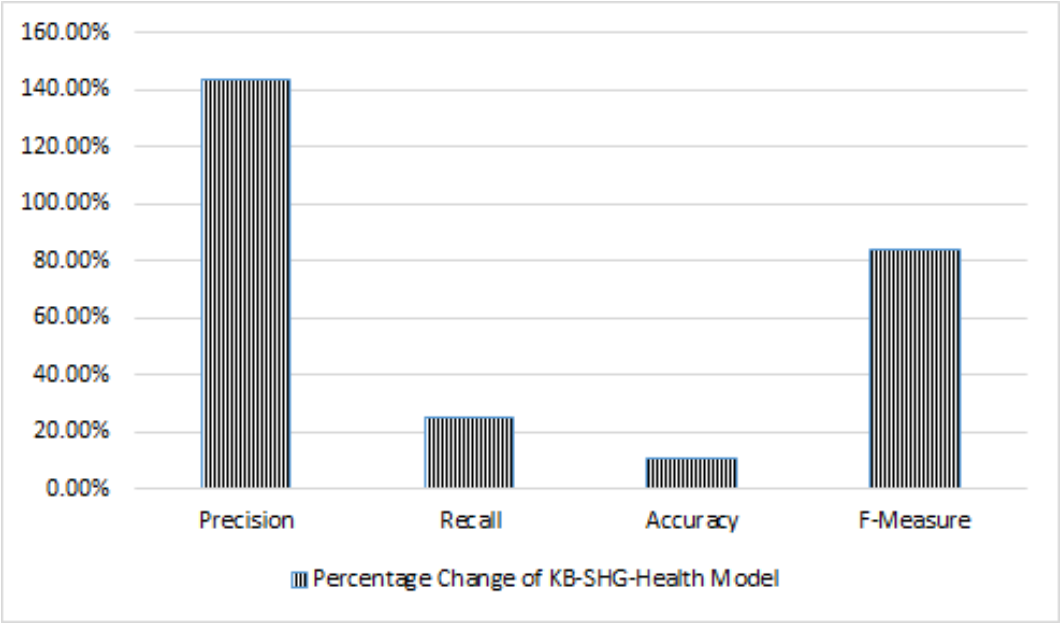
FIGURE 4.8: The improvement of the KB-HIG Model compared to the HIG model

Q1) What is the impact of the knowledge base on the performance of classification model? Applying the knowledge base in the classification model to predict diabetes mellitus helped the KB-HIG model obtain significant improvement for both precision and accuracy as in Table 4.6, Table 4.7, and Figure 4.8. Overall, the performance of the KB-HIG model for both the NHANES and NAMCS dataset raises approximately 15% and round 17% compared to the model that does not apply the knowledge base(HIG model), respectively. This result demonstrated that using the knowledge base could boost the classification model achieving high performance.

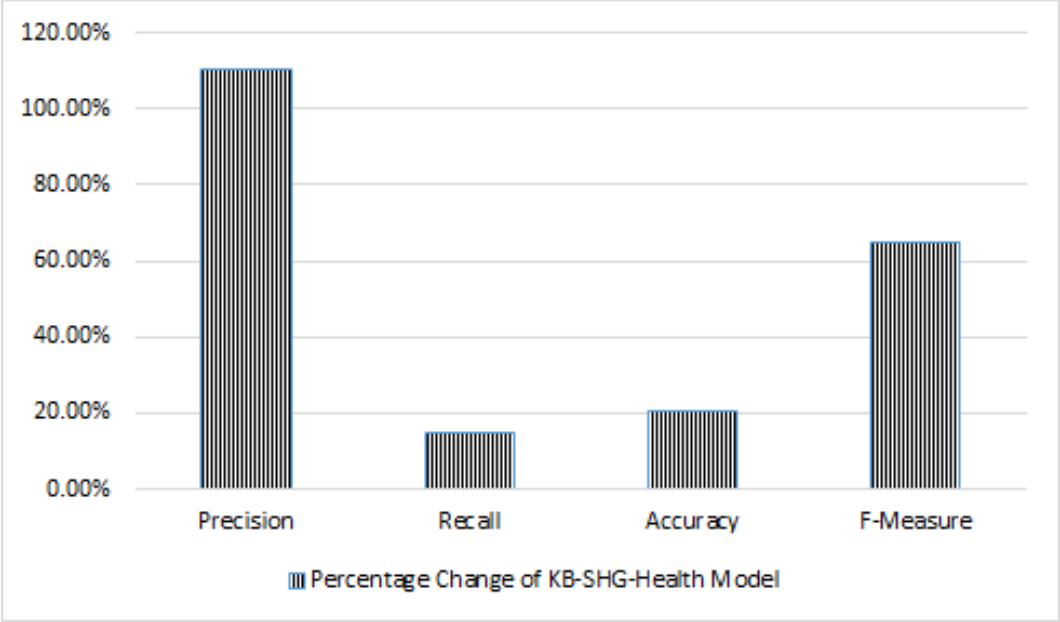
In previous work, the research suggested a new method of building the HIG. Then, the study developed a classification model by using this heterogeneous information graph. The result showed that the HIG model achieved a higher performance than the SHG-Health model. The reason for this advantage was applying the *Pearson correlation* and *semantic relation* to constructing the HIG. The graph helped to obtain an in-depth understanding between the semantic classes, and rejected objects that were not related to the topic. Therefore, the HIG model received a significant improvement in the prediction of health risk status. The HIG model was upgraded by applying knowledge base that was learned from MEDLINE called the KB-HIG model. By using knowledge base in developing the model, the performance of the classification model was significantly improved.

Q2)Does the knowledge base have contributions to other models?

In this study, the knowledge base was integrated on the SHG-Health model to assess the influence of the knowledge base on the SHG-Health model. The SHG-Health model used the information neighbourhood node



(A) NHANES



(B) NAMCS

FIGURE 4.9: The improvement of the KB-SHG-Health Model compared to the SHG-Health model

TABLE 4.8: Comparison between KB-SHG-Health Model and SHG-Health model by NHANES, where the emphasised values indicate the superior performance in comparison.

	Precision	**Recall**	**Accuracy**	***F*-Measure**
KB-HIG model	**0.70066**	**0.68909**	**0.95333**	**0.69422**
SHG-Health model	0.28781	0.54998	0.86052	0.37753
Percentage Change	143.44%	25.29%	10.79%	83.88%

TABLE 4.9: Comparison between KB-SHG-Health Model and SHG-Health model by NAMCS, where the emphasised values indicate the superior performance in comparison.

	Precision	**Recall**	**Accuracy**	***F*-Measure**
KB-HIG model	**0.59257**	**0.65309**	**0.88764**	**0.61954**
SHG-Health model	0.28125	0.56780	0.73726	0.37563
Percentage Change	110.69%	15.02%	20.40%	64.93%

of the HIG to predict the health status. The results from Table 4.8, Table 4.9, and Figure 4.9 show that the model, applying the knowledge base, has a better performance than that of the original model. The performance of the model within the knowledge base achieved a significant improvement with more than 80% from the experiment on the NHANNES dataset and nearly 75% from the experiment on the NAMCS dataset compared to the model without the knowledge base.

The SHG-Health model only considered the HIG constructed based on the information neighbourhood node. The model skipped nodes associated with the topic if these nodes were not neighbourhood nodes, although these nodes were connected to the topic. Moreover, all neighbourhood nodes were used to predict the health risk status, even if these nodes were not related to the topic. In contrast, the study used a large number of articles from MEDLINE to populate the knowledge base for the HIG before building the

SHG-Health model. The method helped to remove the information neighbourhood node that was not associated with the topic. This approach led to the knowledge base SHG-Health model obtaining a better performance of classification than the SHG-Health model.

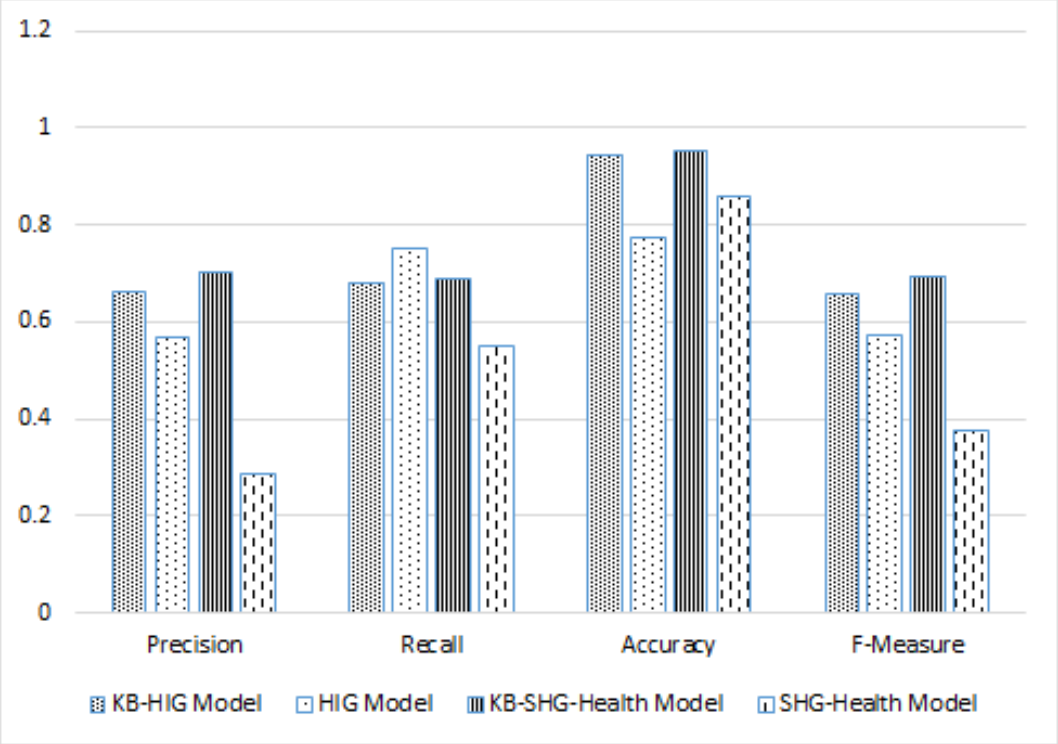


FIGURE 4.10: A comparison of four models by using NHANES dataset

Finally, the study made a comparison among four models which were presented in Figure 4.10 and Figure 4.11. Both models used the knowledge base, including KB-HIG and KB SHG-Health from two datasets, which were better than the HIG model and SHG-Health model without a knowledge base. It was clear that applying the knowledge base onto classification models contributed to the improvement of the classification performance.

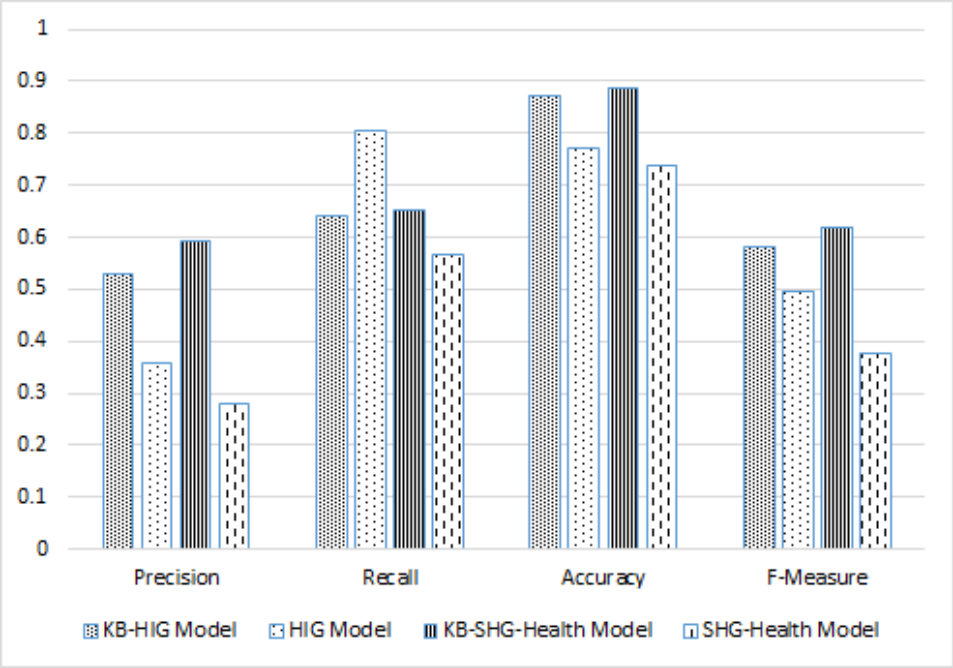


FIGURE 4.11: A comparison of four models by using NAMCS dataset

4.5 Summary

Biomedical literature from MEDLINE has become an important resource to biomedical researchers. In addition, providing decision support systems with high quality is necessary to help practitioners avoid human errors. Therefore, an approach of populating the knowledge base was suggested to improve accuracy in predicting the health risk status by using MEDLINE. After populating the knowledge for the HIG by using instances that were learned from the knowledge base, a classification model was constructed to mine the HIG for predicting the health risk status. The result of the proposal yielded a significant improvement, which advanced the accuracy of the prediction. The research also demonstrated that applying the knowledge base into the classification model has achieved a higher performance than models without applying the knowledge base.

In this Chapter, the study has demonstrated that the performance of the HIG model and SHG-Health has been improved by integrating knowledge-based into graph-based classifiers. The experimental result in Section 4.4.5 showed that the performance of these classification models has been increased significantly. The study has contributed to an innovative framework in integrating knowledge-based into classification models. The proposed framework has helped to improve the accuracy of the patients' health risk prediction as well as to motivate future researchers in using the knowledge base in their model.

Chapter 5

Multi-label Positive and Negative Graph for Predicting Multiple Diseases

Based on the advantages of the knowledge-base graph being built in Chapter 4, this chapter introduces an innovated framework to develop a multi-label classification. This study proposes a separation of the space of a graph into positive and negative graph to discover the relevance of its label through the neighbouring nodes. The method helps to identify essential factors that affect its label. The study has transformed all possible labels for each patient to become a new single combination label set to facilitate the learning of the multi-label classification. The new label set is a subset of the original label set. The approach aims to explore the dependence among different labels. Improving the semantic relationship among labels contributes significant benefits to reduce the ambiguous dependent relationships. Based on the suggestion, a ranking algorithm has been introduced to learn multi-label classifications.

5.1 Introduction

Many researchers have succeeded to learn multi-label classifications in different fields such as image classification (Sun et al. [2014], Luo et al. [2013]), video annotations (Dimou et al. [2009], Nasierding et al. [2015]) and text mining (Zhao et al. [2013], Nam et al. [2014], Tao et al. [2018]), especially in healthcare, it is important to support medical practitioners in predicting multiple diseases for a patient. Building multi-label classification models in diagnosis can support medical practitioners in avoiding human errors. However, learning multi-labels classification is still a challenging task because of label ambiguity and data complexity (Zhang et al. [2018b]), uncertain data (Liu et al. [2016]), and order of sample similarities (Cai and Zhu [2017]).

In the last decade, researchers have used three approaches: problem transformation, algorithm adaptation, and the ensemble method to address multi-label classification. The technique of problem transformation is known as binary relevance and label powersets, which transforms the multi-label classification problem into one or several single-label classification problems. As another example, the method of algorithm adaptation adjusts traditional single-label classification algorithms that deal directly with multi-label classification (Chen et al. [2016b]). Finally, the ensemble method is one of the most flexible approaches as it has a significant effect on multi-label learning because of its purposed combination of all state-of-the-art techniques to process learning. By applying this method to multi-labels, Ringsquandl et al. [2016] constructed a knowledge graph based on mining discriminative sub-graph patterns to explore the label correlations through the extractive link among labels. This framework derives from previous

work in Wu et al. [2014], which used a bag of graphs labelled as positive or negative examples to learn a multi-graph classification. The problem of graph classification has also been studied by Kong and Philip [2012], who presented an algorithm to learn multi-label feature selection through optimal subgraph features. By using sub-graph patterns with constraints, they showed that the search space can be pruned, and extraction of consistent patterns can be guaranteed. The approach has helped to boost the accuracy of classification.

To solve multi-label classification, many researchers focused on dealing with learning label correlation (Xu et al. [2016], Huang et al. [2016], Zhang et al. [2018a], Han et al. [2019]), which plays an important role in learning multi-label classification. They tried to minimise label-specific features and facilitate label-specific data representation to obtain high-performance classifications. The concepts underlying this method were to decrease the computational cost for each label by their relevance. Then, labels with less effective contributions to others were rejected to help improve accuracy by cooperatively identifying the accurate classifiers. In addition, some researchers paid attention to ambiguity among classes or missing data by Ma and Chow [2019], Zhu et al. [2018], Huang et al. [2019] to boost the performance of multi-label classification. However, semantic relations among labels have not been thoroughly exploited through all labels of the training dataset.

In healthcare, a patient might suffer from diabetes and liver cancer simultaneously. Some researchers have built models to support medical practitioners in medical diagnosis (Li et al. [2016, 2017], Zhang et al. [2019]). These models were developed by using electric health records for predicting multiple diseases. However, their relevance to diseases has not been

paid sufficient attention to develop multi-label classification models. Discovering links betwwen diseases and determining relationships between symptoms and diseases plays an important role in diagnosis. Therefore, the predicted results are more accurate if classification models have the capacity to discover these kinds of relationships. In addition, literature reviews of medical sources to evaluate the accuracy of predicted results have not been utilised. Developing classification models based on health examination records might not ensure the accuracy of predicted results. It may be hard to avoid human errors if the model is only based on experiment-based to give predicted results.

5.2 Research Problem Formulation

Given a patient space $\mathcal{P} = \{p_1, p_2, ..., p_n\}$ as defined in Definition 3.1, and a disease space $\mathbf{D} = \{d_1, d_2, ..., d_m\}$ as defined in Definition 4.7. It can be assumed that an instance $p_i \in \mathcal{P}$ is associated with a subset of labels $\mathcal{D}, \mathcal{D} \subseteq \mathbf{D}, |\mathcal{D}| = 2^m$. $\mathcal{F} = \{(p_i, \mathcal{D}_i)\}_{i=1}^{N}$ is denoted as the set of training instances. The goal of multi-label classification is to build a function f(p) that maps an instance p_i to its associated set of labels $\mathcal{D}_i$.

In a medical diagnosis, in Table 5.1, a patient suffers from diabetes and heart disease simultaneously. This type of data can be used to learn a multi-label classification task. In this case, each patient might be linked to multiple diseases. In real life, different conditions may have the same or different effects on a similar set of attributes. For example, if a patient suffers from liver and kidney diseases, the problem may be caused by habits such as high weekly alcohol intake. However, habits like smoking ten cigarettes

per day may affect only the liver. This study not only identifies how many diseases the patient suffers but also explains different effects among properties. These attributes guarantee a significant contribution toward classifying multiple diseases.

TABLE 5.1: Example of multiple diseases in a real dataset

Patients	**diabetes**	**heart disease**	**asthma**
p_1	x		x
p_2			x
p_3	x	x	

Assume that a patient p_i belongs to space P, P $\in \mathcal{P}$, which is represented by multiple features. A patient can be characterised by their name, weight, age, use of alcohol, and smoking habits. These attributes can be categorised into different health topic types such as profile, disease, habit. A patient can also be presented as a vector: $p =< a_1, \ldots, a_h >$, where h is the number of features that identify the dimension of the input space, and a_i is the value of the i-th feature. The value of this feature can be numeric feature or categorical. Given the output space D $= \{d_1, d_2, \ldots, d_m\}$ is a set of disease labels , where m is the number of all possible disease labels, D $\subseteq$ **D**. Given also P $= \{p_1, p_2, \ldots, p_n\}$ is a set of n patients, P $\subseteq \mathcal{P}$. Each patient in P is associated with one or more disease labels in D. The task of this research is to build a model which predicts that a disease subset belonging to D is associated with each patient.

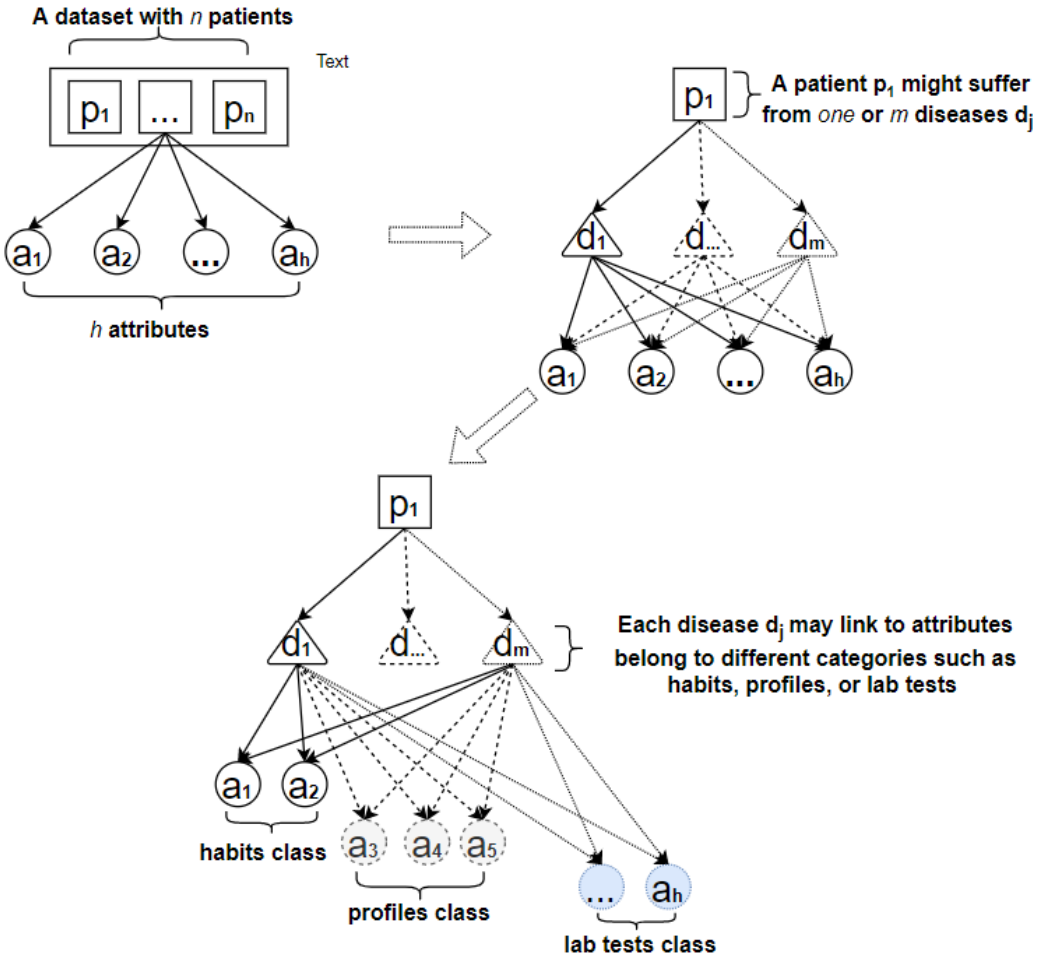


FIGURE 5.1: Presentation of an information network for a dataset. A patient p_1 may suffer one of m diseases $d_i, (i = 1, 2, ...m)$. Each disease d_i may be assigned to h attributes belong to different semantic groups such as habits, profiles, or lap tests.

5.3 Framework

In medical diagnosis, learning classification requires extensive computational resources. The cause and effect relationship plays a vital role in evaluating the probability of a diagnosis (Rebitschek et al. [2016]). Exploiting these kinds of relationships is not an easy task in automatic classification learning. Therefore, this research explores the effect of different attributes, such as alcohol use or smoking, leading to various diseases. In fact, each patient may have one or more diseases associated with a list of attributes. These causes could belong to different categories such as profile, habit, lab test. These data may be presented as information networks, as shown in Figure 5.1.

The study treats attributes of observational data as nodes in network information. These attributes could belong to the same or different objects; therefore, the links between these nodes are different. These attributes contain target labels as well as features that may have directly or indirectly affected target labels. Therefore, these contributes are a part of a heterogeneous information graph $G :=< V, E, M >$, where V is a set of nodes, $V = \bigcup_i^t \mathcal{A}_i$, $\mathcal{A}_i = \{a_{i1}, a_{i2}, \ldots, a_{ih}\}$, $i = 1, 2, ..., t$ are t types of data objects and $t \geq 2$, E and M is the set of links between any two objects V and the set of weight values by links, respectively.

A heterogeneous information graph is one of the most effective approaches for dealing with multi-label learning, (Tahir et al. [2012], Kong et al. [2013], Zhou and Liu [2014], Dos Santos et al. [2016]). Applying a heterogeneous information graph is advantageous in exploiting relationships among different types of entities. Mining the linkage structure of a heterogeneous

information graph could help to deal with sample imbalance and label correlation. Taking advantage of these methods, this study presents a heterogeneous information graph G as a multi-label graph $\mathcal{G}$.

Definition 5. 1. ***[Multi-label Graph]***

The multi-label graph is a 4-tuple: $\mathcal{G} =< V, E, \mathcal{L}, \mathcal{Y} >$. *V and E is a set of nodes and edges, respectively.* $\mathcal{L} = \{l_1, l_2, ..., l_m\}$ *is a set of m possible labels in the graph.* $\mathcal{Y}_i = (y_i^1, y_i^2, ..., y_i^m) \in \{0, 1\}^m$ *specifies multiple labels that are linked to the graph* $\mathcal{G}$. *If* y_i^k *equals 1, the k-th class label will be assigned to graph* $\mathcal{G}$. *If* y_i^k *equals 0, the k-th class label will not be assigned to graph* $\mathcal{G}$. □

The study tries to learn all subgraphs of a multi-label graph $\mathcal{G}$. Each subgraph indicates one of the corresponding graph types as a single class label. Therefore, this study considers an instance x_i as a graphic instance, which can be presented by a vector $\overrightarrow{x_i} = [x_i^{g_1}, x_i^{g_2}, ..., x_i^{g_m}], \{g_1, g_2, ..., g_m\}$ as a set of subgraphs belonging to $\mathcal{G}$. Each subgraph is one of the target labels assigned with a graphic instance x_i (A graphic instance is an example of m subgraph).

This study considers both positive and negative subgraph transformations. A technique, which has succeeded in the learning of label correlation has helped boost classification effectively (Huang et al. [2017b], Wu et al. [2014]). There are m subgraphs that belong to a graphic instance $\mathcal{G}$. Therefore, there is a 2^m set of multi-label graph $\mathcal{G}_{x_i}$. The goal of this study is not only to mine cooperation among subgraphs but also to exploit their effectiveness to deal with the learning multi-label graph problem. A rank coefficient for each subgraph is computed for contribution on learning label correlation. The study also separates a subgraph into two spaces: one is for the

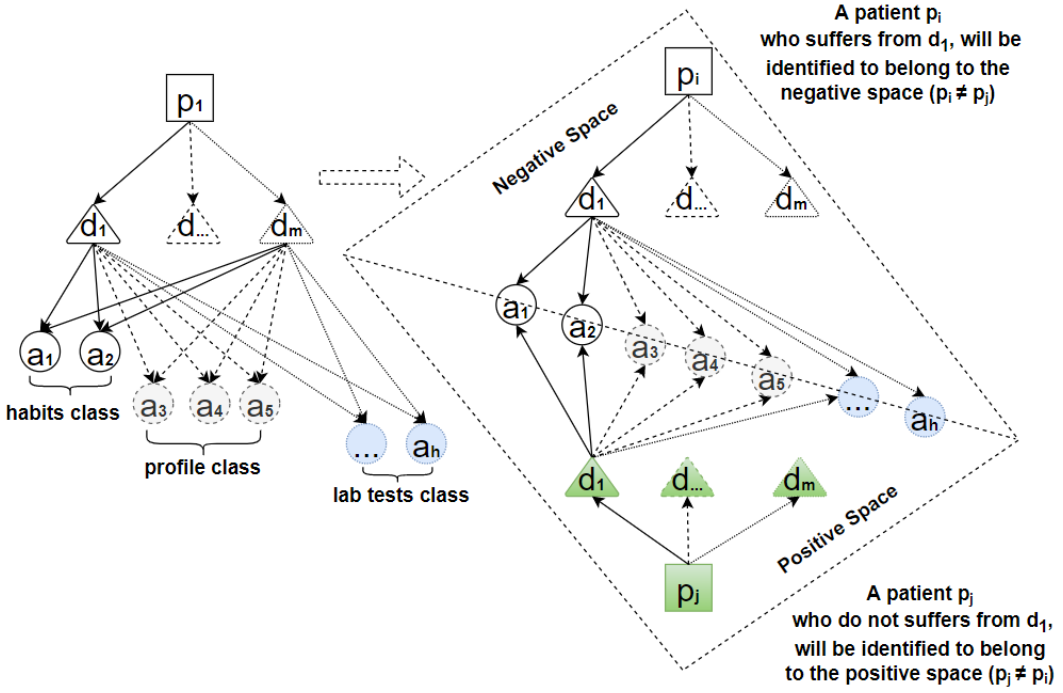


FIGURE 5.2: A heterogeneous information graph is divided into positive and negative space. If a patient p_1 suffer from one of m diseases d_i, $(i = 1, 2, ...m)$, this patient will belong to the negative graphic space. A patient will belong to positive graphic space if they do not suffer from any diseases d_i.

positive subgraph, and one is for the negative subgraph. The positive subgraph indicates a true class label for a graphic instance $x_i^{g_k}$ and the negative subgraph indicates a false class label for a graphic instance $x_i^{g_k}$. Figure 5.2 indicates a two space of a graphic instance $x_i^{g_k}$. In this case, x_i represents for both p_i and p_j where g_k represents for disease class d_1. p_i belong to the negative space because p_i suffers from d_1 and p_j belong to the positive space because p_j does not suffer from d_1, $p_i \neq p_j$. One of tasks in this study is to identify the value for both patient assigned within d_1 and without d_1.

5.3.1 Building a Knowledge Graph

A multi-label graph $\mathcal{G}$ is a map of m single graph G. To construct a graph G, the research follows the ***Definition 3.6*** to build a knowledge graph G. $G :=< V, E, M >$, with an object mapping function $\varphi : V \rightarrow A$ and a link type mapping function $\psi : E \rightarrow R$, where M is a matrix $A \times A$; each object $v_i \in V$ belongs to one particular class in the semantic class set $A : \varphi(v) \in A$ and $A = \{a_1, a_2, \ldots, a_d\}$ is a set of attributes for each patient in the real dataset; each link $e_j \in E$ belongs to a particular relation type set in the relation type set $R : \phi(e) \in R$, where $|R| = 8$, eight medical semantic classes as in Table 3.2 and e is the link connecting two objects defined by a *Pearson correlation* coefficient ρ. The value of weight among a pairwise $\rho(v_i, v_j)$ is computed by Eq.(3.1). Using Pearson Correlation and Sematic Relations, researchers have achieved significant effectiveness on learning classifications based on the graph (Pham et al. [2018]). After building a graph G, a mapping function $f(\mathcal{Y}_i) : \mathcal{L} \rightarrow \mathcal{A}$ is conducted to generate a multi-label graph $\mathcal{G}$.

5.3.2 Population of Knowledge Graph

In a multi-label graph, each node belongs to one or more class labels. The multiplicity depends on the mapping function $f(\mathcal{Y}_i)$. The study divides the multi-label graph into many subgraphs, which belong to only one class label. The idea of the method is to consider each node of the graph as a concept, which represents a set of terms being mined from the categoric data. Based on the previous work (Pham et al. [2019]), the study populates the medical knowledge for graph $\mathcal{G}$ by mining MEDLINE[1], a corpus with the citations of publishing papers in the medical domain. The approach

[1]https://www.nlm.nih.gov/bsd/medline.html

helps to discover more hidden meanings among two nodes. The medical knowledge $\mathbb{MK}$ could be defined as follows:

Definition 5. 2. ***[Medical Knowledge]***
$\mathbb{MK}$ *is a 3-tuple* $\{\mathfrak{D}, \mathcal{L}, \mathcal{T}_{\mathfrak{D}}^{\mathcal{L}}\}$. $\mathfrak{D} = \{\mathfrak{d}_1, \mathfrak{d}_2, \ldots, \mathfrak{d}_z\}$ *is a set of documents.* $\mathcal{L}$ *is a set of the possible topic or possible class labels that links to* $\mathfrak{D}$. $\mathcal{T} = \{t_1, t_2, \ldots, t_q\}$ *is set of terms or concepts found by mapping function* $\mathcal{L} \longrightarrow \mathfrak{D}$. □

The value Ψ for each term t is computed throughout the word2vec algorithm suggested by Mikolov et al. [2013]. Every single term is assigned a corresponding vector which indicates a specific weight. The ranges value Ψ for each term is from zero to one. Obtaining the weight of term can help to populate the knowledge for the graph $\mathcal{G}$. By integrating the medical knowledge for the graph, the multi-label graph was indicated as $\mathcal{G} =< V, E, \mathcal{L}, G_{\mathcal{Y}}^{\mathbb{MK}} >$. $\mathbb{MK}$ is a set of documents for each topic belonging to $\mathcal{Y}$. Each topic corresponds to a single subgraph or single class label. $|G_{\mathcal{Y}}^{\mathbb{MK}}| = m$, m is the possible class label of a graphic instance, given $\acute{G} = \{\acute{g_1}, \acute{g_2}, ..., \acute{g_m}\}$ is a set of single label graphs. Each single label graph is identified by function $f(\acute{g})$.

$$f(\acute{g}) = \sum_{i=1}^{k} v_i * \rho(g, v_i) * |\langle \mathcal{M}_{\mathcal{A}}^{\mathcal{T}}, \overrightarrow{\mathcal{M}_{\mathcal{A}}^{\mathcal{T}}} \rangle| = \sum_{i=1}^{k} v_i * \rho(g, v_i) * \Psi(g, (a_i, t_i)) \quad (5.1)$$

where v_i is a vertex for a graphic instance g, $v_i \in V, V \subseteq \mathcal{A}$. $\rho(g, v_i)$ is weight between vertex v_i and labelled vertex of an graphic instance g. $\mathcal{M}_{\mathcal{T}}^{\mathcal{A}}$ is a matrix $\mathcal{A} * \mathcal{T}$. For each pair $(a_i, t_i) \in \mathcal{M}_{\mathcal{T}}^{\mathcal{A}}$, the value of weight $(a_i, t_i) \in \overrightarrow{\mathcal{M}_{\mathcal{A}}^{\mathcal{T}}}$ indicate exist of a map between a_i and t_i. $\Psi(g, (a_i, t_i))$ is the value

learned from medical knowledge, which is mapped for a graphic instance g.

5.3.3 Possible and Negative Graph

With the separation of the multi-label graph into two spaces, $\mathcal{G}$ can be a set of positive and negative single graphs, $\mathcal{G} = \{< g_1^+, g_1^- >, < g_2^+, g_2^- > , ..., < g_m^+, g_m^- >\}$. In this case, this study not only can exploit the correlation among single graphs (single label) but also can mine the effectiveness of graphs among two spaces (positive and negative). There are differences relating to the coefficient among different subgraphs and between positive and negative subgraphs. All single label graphs are calculated by Eq.(5.1). A rank value is used to identify the effect among positive and negative subgraphs through an average value, which is the following:

$$\arg(\mathcal{G}^{l_j}|g_i^-) \leq \delta_{(l_j)} \leq \arg(\mathcal{G}^{l_j}|g_i^+), where\ \delta \in \Delta,\ l \in \mathcal{L}, |\mathcal{L}| = m \tag{5.2}$$

m is possible single label class from training dataset, $|g_i^-|$ and $|g_i^-|$ are graphic instances for positive and negative from a training dataset. $\Delta = \{\delta_1, \delta_2, ..., \delta_m\}$ indicates m-th rank values corresponding to m single label graph.

5.3.4 Learning Multi-label Classification

In this study, the number label of a multi-label graph was predicted based on both relevance among single label graphs and through positive and negative space. Therefore, the score function was used to identify the ranking

value for every single graph by Eq.(5.2). Given a set of corresponding values Δ to all single label graphs, the values for each $\delta_j \in \Delta$ can be presented as:

$$\min(\delta_j) \leq \delta_j \leq \max(\delta_j) \tag{5.3}$$

where $\min(\delta_j)$ corresponds to as a false label assigned to a graph, and $\max(\delta_j)$ is indicated as a true label assigned to a graph.

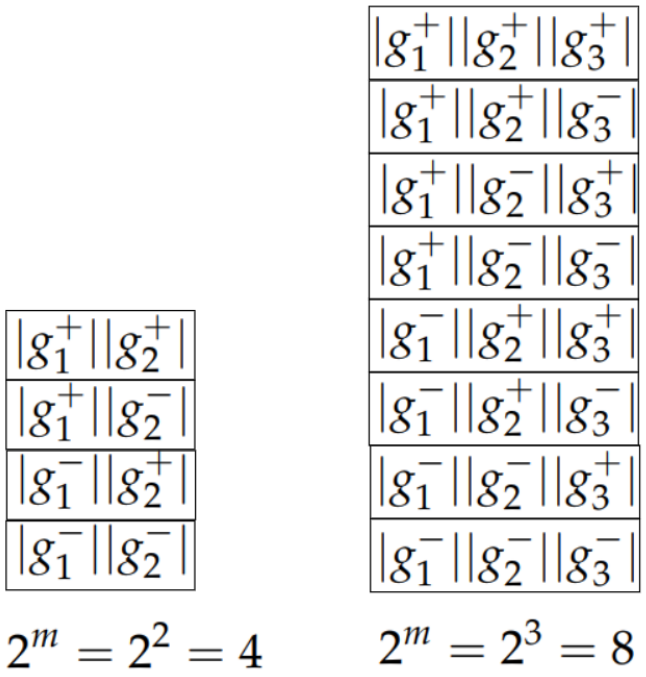


FIGURE 5.3: An example of combinational labels for a training set within two and three original labels. If the training set has two labels, the number of combinational labels is four. The number of combinational labels will increase if the number of labels in training set increase.

Based on separating a graphic instance into two space instances, the study used the label powerset method (Abdallah et al. [2016]), which transformed m-th multi-label graph classification into classifying 2^m combinational labels. All value of each graphic instances were computed by Eq.(5.2)

and Eq.(5.3). After identifying the value for each graphic instance, a matrix is defined to specify the cooperation between graphic instances. Figure 5.3 shows an example of cooperation among graphic instances for a training set with two and three original lebels. Each new combinational label set $\acute{\mathcal{L}}$ has a set of original single label sets from $\mathcal{L}$. The greater the number of the original single label sets is, the more the number of combinational labels is. The new label set $\acute{\mathcal{L}}$ is treated as single class label that needs to be predicted. The function $\mathbf{CL}(\acute{\mathcal{L}})$ is to identify values for each new combinational label set, which is determined as follows:

$$\mathbf{CL}(\acute{\mathcal{L}}_i) = \sum_{j=1}^{m}(\min(\delta_j) \mid \max(\delta_j)), \delta \in \Delta, i = 2^m \tag{5.4}$$

where $i = 2^m$, the number of combinational labels. After calculating all new combinational labels based on Eq.(5.4), a set of combinational labelled values $\mathbf{LB} = \{\mathbf{CL}(\acute{\mathcal{L}}_1), \mathbf{CL}(\acute{\mathcal{L}}_2), ..., \mathbf{CL}(\acute{\mathcal{L}}_{2^m})\}$ was generated. To determine cases and values of combinational labels, Algorithm 5.1 is presented.

Algorithm 5.1 Algorithm for identifying the combinational labels.

1: **INPUT**: Given a set number of original label $\mathcal{L} = \{< \min(\delta_{l_1}), \max(\delta_{l_1}) >, < \min(\delta_{l_2}), \max(\delta_{l_2}) >, ..., < \min(\delta_{l_m}), \max(\delta_{l_m}) >\}$.
2: **OUTPUT**: A list of combinational labelled values **LB**.
3: **for** $i \leftarrow 0$ to 2^m **do**
4: $binary = 1$;
5: **for** $j \leftarrow 0$ to m **do**
6: **if** $((i \,\&\, binary) > 0)$ **then**
7: $\mathbf{CL_i}(\acute{\mathcal{L}}_i) = \mathbf{CL_i}(\acute{\mathcal{L}}_i) + \max(\delta_j)$;
8: **else**
9: $\mathbf{CL_i}(\acute{\mathcal{L}}_i) = \mathbf{CL_i}(\acute{\mathcal{L}}_i) + \min(\delta_j)$;
10: **end if**
11: binary = binary « 1;
12: **end for**
13: **end for**
14: **return** $\mathbf{CL_i}(\acute{\mathcal{L}}_i) \in \mathbf{LB}$

The input of the Algorithm 5.1 is a set of m original labels and the output is 2^m set of combinational labelled values. Each original label has been assigned with two values (min and max), which are identified by Eq.(5.2) and (5.3). The algorithm first uses a for loop run from 0 to 2^m. Then, a value bit *'binary'* in binary representation equals 1. Next, another loop is used to check m original labels. Then, the algorithm checks the bit value of i, $i = 1, 2, ..., 2^m$ in binary representation. If the bit value is greater than 0, the max value of *m-th* is added into the set $\mathbf{CL_i}(\acute{\mathcal{L}}_i)$, otherwise, the min value of *m-th* is added into the set $\mathbf{CL_i}(\acute{\mathcal{L}}_i)$. Then, after checking the next bit, i is run. Finally, the algorithm returns a set of combinational labelled values **LB**. The Algorithm 5.1 uses two nested loops to identify a set of **LB**. The complexity of the Algorithm 5.1 is identified by $O(n^2)$.

assume $\mathbb{L} = \{\acute{\mathcal{L}}_1, \acute{\mathcal{L}}_2, ..., \acute{\mathcal{L}}_{2^m}\}$ was a set of new combinational labels. The task of this study was to predict a combinational label $\acute{\mathcal{L}}$ for a graphic instance $\mathcal{G}$. The study introduced Algorithm 5.2 called **Mul**ti-label **G**raph **Ra**nking **L**earning (**MulGRaL**) algorithm to deal with the multi-label classification. The algorithm first uses two nested loops to rearrange the values of the combined labels **LB** in descending order. Each combinational value $\mathbf{CL}(\acute{\mathcal{L}}_i) \in \mathbf{LB}$ is calculated by Eq.(5.4). Then, using another loop is to create a new list of values belonged to $\mathbf{L\acute{B}}$. Each new value is identified by averaging between two contiguous values of **LB** sorted in descending order. Then, a for loop is used to check all values of $\mathbf{L\acute{B}}$. The model is eventually run to predict labels associated with a graph by combining them with a threshold value (α) to optimise the algorithm results.

Algorithm 5.2 Algorithm for Multi-label Graph Classification.

1: **INPUT**: Given a set number of combine labels $\mathbb{L}$ corresponding to a set of combinational labelled values **LB**.

- $\mathbb{L} = \{\acute{\mathcal{L}}_1, \acute{\mathcal{L}}_2, ..., \acute{\mathcal{L}}_{2^m}\}$
- $\mathbf{LB} = \{\mathbf{CL}(\acute{\mathcal{L}}_1), \mathbf{CL}(\acute{\mathcal{L}}_2), ..., \mathbf{CL}(\acute{\mathcal{L}_{2^m}})\}$
- Given a graphic instance $\mathcal{G}_i$.

2: **OUTPUT**: an associated between a combine label $\acute{\mathcal{L}}_j$ and a graphic instance $\mathcal{G}_i$.
3: **Initialise** $|\mathbf{LB}| = 2^m, temp = \mathbf{CL}(\acute{\mathcal{L}}_1)$;
4: **for** $i \leftarrow 1$ to $(2^m - 1)$ **do**
5: **for** $k \leftarrow i + 1$ to 2^m **do**
6: **if** $\mathbf{CL}(\acute{\mathcal{L}}_i) <= \mathbf{CL}(\acute{\mathcal{L}}_k)$ **then**
7: $temp == \mathbf{CL}(\acute{\mathcal{L}}_i)$
8: $\mathbf{CL}(\acute{\mathcal{L}}_i) == \mathbf{CL}(\acute{\mathcal{L}}_k)$
9: $\mathbf{CL}(\acute{\mathcal{L}}_k) == temp$
10: **end if**
11: **end for**
12: **end for**
13: **for** $i \leftarrow 1$ to $(2^m - 1)$ **do**
14: $\acute{\mathbf{CL}}(\acute{\mathcal{L}}_i) = (\mathbf{CL}(\acute{\mathcal{L}}_i) + \mathbf{CL}(\acute{\mathcal{L}}_{i+1}))/2$
15: **end for**
16: **return** $\acute{\mathbf{LB}}$.
17: **for all** $\acute{\mathbf{CL}}(\acute{\mathcal{L}}_j) \in \acute{\mathbf{LB}}$ **do**
18: **if** $|\mathcal{G}_i| > (\acute{\mathbf{CL}}(\acute{\mathcal{L}}_1) + \alpha)$ **then**
19: $|\mathcal{G}_i| == \acute{\mathbf{CL}}(\acute{\mathcal{L}}_1)$;
20: **else**
21: **if** $|\mathcal{G}_i| <= (\acute{\mathbf{CL}}(\acute{\mathcal{L}}_j) - \alpha)$ and $|\mathcal{G}_i| > (\acute{\mathbf{CL}}(\acute{\mathcal{L}}_{j+1}) + \alpha)$ **then**
22: $|\mathcal{G}_i| == \acute{\mathbf{CL}}(\acute{\mathcal{L}}_{j+1})$;
23: **else**
24: **if** $|\mathcal{G}_i| < (\acute{\mathbf{CL}}(\acute{\mathcal{L}}_{(2^m-1)}) - \alpha)$ **then**
25: $|\mathcal{G}_i| == \acute{\mathbf{CL}}(\acute{\mathcal{L}}_{2^m})$;
26: **end if**
27: **end if**
28: **end if**
29: **end for**
30: **return** $\acute{\mathcal{L}}_j$

The complexity of the Algorithm 5.2 is identified by $O(n^2)$. The Algorithm 5.2 firstly uses two nested loops to arrange an array in ascending order. Later, the algorithm uses another loop to generate new values for $\mathbf{L\acute{B}}$. The algorithm finally uses one more loop to find the target combination label. Overall, the complexity is identified for the Algorithm 5.2 by $O(n^2)$ + $O(n)$ + $O(n)$, which equals $O(n^2)$.

5.4 Experiments and Evaluation

5.4.1 Experimental Design

Firstly, a heterogeneous information graph is adopted into the observation data to ascertain the effectiveness among objects in the National Health and Nutrition Examination Survey (NHANES)[2] dataset. NHANES is a dataset which collects patient information including profile, lab test or habits. Using a graph for multi-disease learning could help to infer cause-effect relationships. These kinds of relationships play essential roles in diagnosis. A patient may have some diseases and the diseases may cause various symptoms. However, the same symptoms may cause multiple conditions. Therefore, hidden relationships may be exploited through a graph. The approach is advantageous for multi-label learning because it helps to facilitate the relevance of mining labels. Secondly, this study enriches the knowledge for the graph by instances of learning from MEDLINE, which contains the citations of articles published in the medical domain. The method can help to remove non-effective relationships and improve the accuracy of learning. Increasingly enriching the graph by integrating more knowledge helps

[2]https://www.cdc.gov/nchs/nhanes/index.htm

the reliability of classification models. These models, therefore, are more suitable to be used to support practitioners in diagnosis. Finally, a ranking algorithm based on the framework of the multi-label graph is proposed to deal with the multi-label classification problem.

In this study, the NHANES dataset was used for the experiment. The dataset was randomly divided into five subsets as a way to adopt the *5-fold* approach (Yadav and Shukla [2016]) to help ensure the reliability of evaluation results. Each subset had a training set and a testing set. The former was used to help develop the multi-label classification model, whereas the latter was used for evaluation of the proposed model. Each subset corresponded to each round of the experiment. An average of five rounds was used for the final result of the test. Experimental results of the proposed model were compared to the baseline model to show the percentage of improvement.

5.4.2 Dataset

A multi-disease dataset NHANES was used to run the experiment. Twelve diseases from the dataset were used. Figure 5.4 presents the number of patients for each disease. NHANES comprises a series of surveys including lab tests, physical examinations and personal habits in different population groups. The dataset has more than 2585 attributes. However, only 318 attributes were considered for the experiment after cleansing data by removing noise data. This research only considered attributes with less than 50% of the missing data. Using all the characteristics in this study led to a high percentage of missing data, resulting in low accuracy for the model's prediction. Table 5.2 presents the twelve diseases used for the experiment

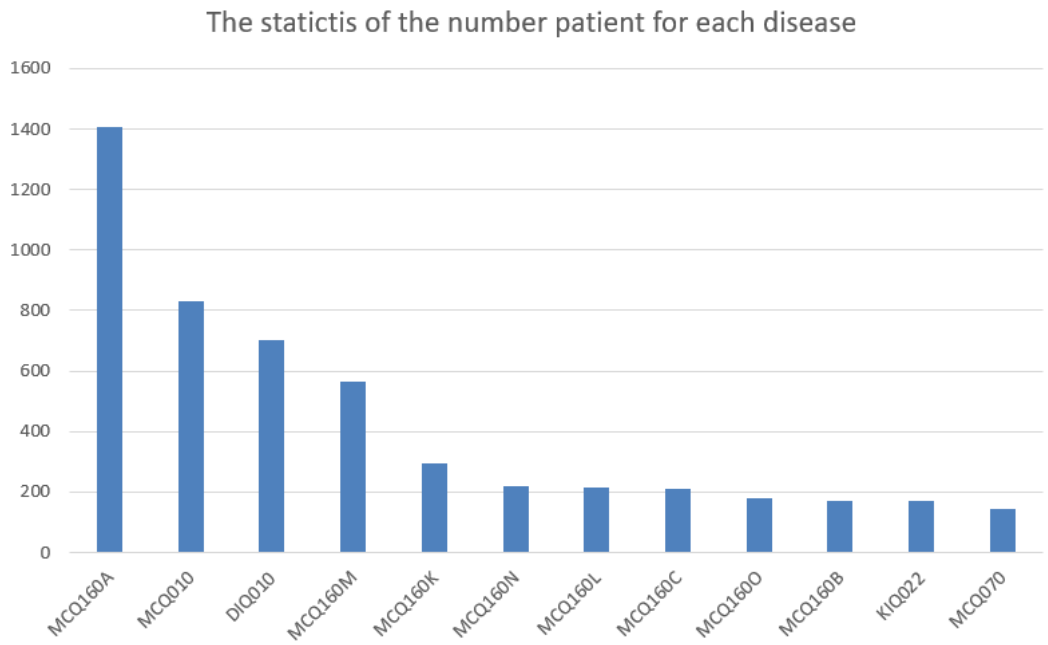


FIGURE 5.4: The number of patients for each disease

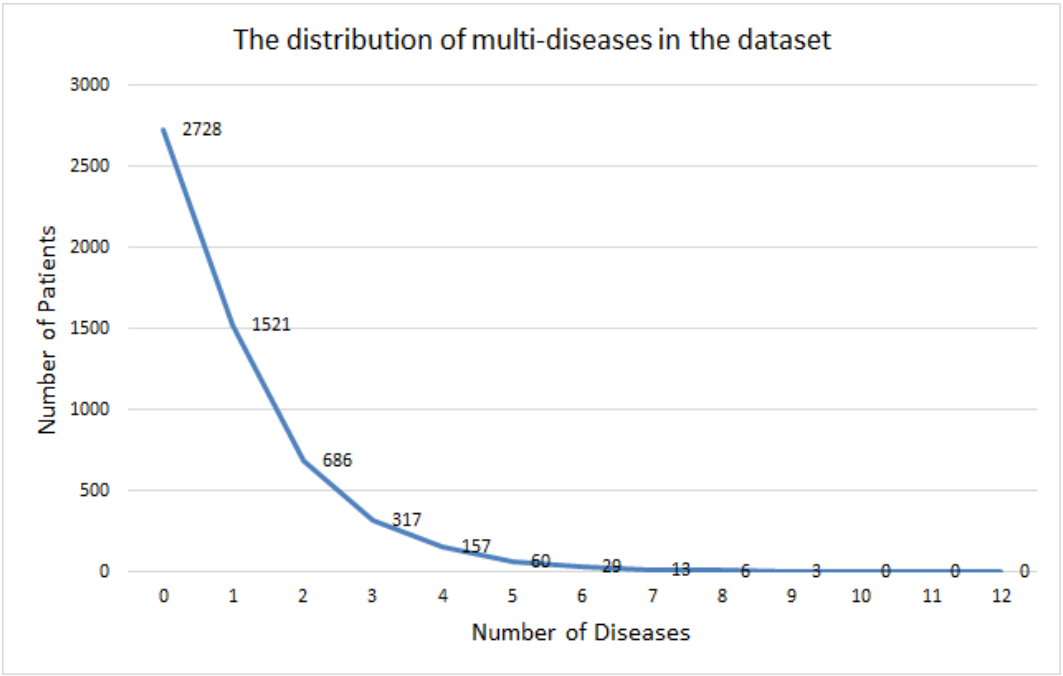


FIGURE 5.5: The number of patients for multi-diseases

Because of the uneven distribution of the number of patients for each

TABLE 5.2: Twelve diseases used for the experiment

Code	Diseases
MCQ160A	arthritis
MCQ010	asthma
DIQ010	diabetes or sugar diabetes
MCQ160M	thyroid problem
MCQ160K	chronic bronchitis
MCQ160N	gout
MCQ160L	liver condition
MCQ160C	coronary heart disease
MCQ160O	COPD
MCQ160B	congestive heart failure
KIQ022	weak or failing kidneys
MCQ070	psoriasis

disease, some classes had overwhelming samples while others had only limited samples as in Figure 5.5. Moreover, the obstacle of data loss occupied a large proportion in this dataset. A preprocessing methods was needed to clean the data. Before proceeding with data cleaning, data transformation was also conducted to handle the conversion of data in a specific format by the binary value. After the preprocessing step, this study analysed and evaluated the data to generate experimental data batches. Figure 5.5 indicates the number of patients for multi-diseases in the NHANES dataset. The percentage of patients with more than six diseases was significantly low and almost zero for patients suffering from ten to twelve diseases. Therefore, in this experiment, a three-level disease hierarchy including layers of five diseases, eight diseases and twelve diseases were generated to compare the effect of each disease layer to the performance of the classification model. Figure 5.6 shows the three-level disease hierarchy, which was designed for the experiment. Each disease layer is a subset of all diseases of the dataset.

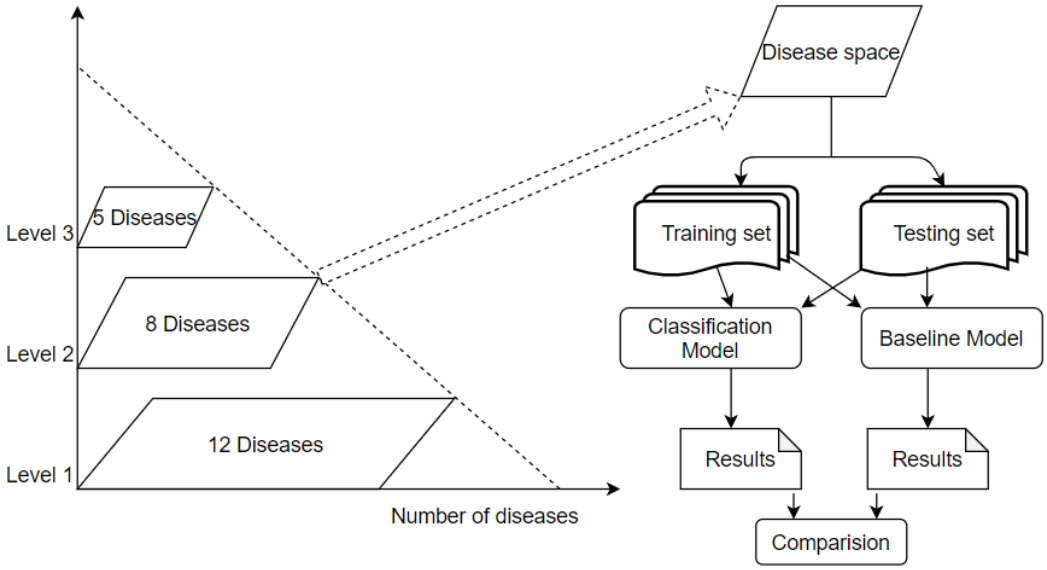


FIGURE 5.6: The three-level disease hierarchy is designed for experiment

5.4.3 Baseline Model

The first baseline model was set up based on Chen's Model (Chen et al. [2016a]). A classification model was built from the heterogeneous information graph within neighbourhood nodes. This study adapted the model corresponding to the semi-supervised heterogeneous graph on health (SHG-Health) algorithm called Mul-SHG-Health model to make predictions of multiple diseases. The approach calls for the problem transformation of dealing with the multi-label classification into binary classification. Problem transformation is one of the methods that is used by researchers, such as Zhang and Zhang [2010], Alvares-Cherman et al. [2012], to deal with the multi-label classification.

Another baseline model (Wang et al. [2020]) was developed based on

Wang's Model. They combined a directed disease network and recommendation system techniques for enhancing the prediction model for multiple disease risks. They first built a directed network based on leveraging chronological orders among consecutive hospital visits. This approach aims to explore temporal relations among diseases. Then, they calculate the risk score of the temporal relations between a disease m and each subset of the target patient's medical history. The risk score of developing disease m is the aggregating risk scores of all the subsets which called an aggregated disease temporal links approach (ADTLM).

5.4.4 Performance Measure

This study used eight multi-label performance measures (Wu and Zhou [2017], Santos et al. [2011]) to evaluate the predicted result of the proposed model. The formulas are as follows below:

Subset accuracy (Santos et al. [2011]) is a metric to identify the exact match between samples and labels classified correctly.

$$Subset - Accuracy = \frac{1}{N}\sum_{i=1}^{N}|y_i = z_i| \tag{5.5}$$

Hamming loss ((Wu and Zhou [2017])) indicates the ratio between the wrong labels and total number labels. The best performance is equal to 0

$$Hamming - loss = \frac{1}{N\mathcal{L}}\sum_{i=1}^{N}\sum_{j=1}^{\mathcal{L}}|y_{ij} \neq z_{ij}| \tag{5.6}$$

Micro-averaged Precision (Santos et al. [2011]) is defined by using the sum-up of all true positives, and false positives on different sets to calculate

the precision.

$$Micro - Pre = \frac{\sum_{l=1}^{\mathcal{L}} TP_l}{\sum_{l=1}^{\mathcal{L}} TP_l + FP_l} \tag{5.7}$$

Micro-averaged Recall (Santos et al. [2011]) is identified by using the sum-up of all true positives, and false negatives on different sets to calculate the precision.

$$Micro - Rec = \frac{\sum_{l=1}^{\mathcal{L}} TP_l}{\sum_{l=1}^{\mathcal{L}} TP_l + FN_l} \tag{5.8}$$

Micro-averaged F-measure (Santos et al. [2011]) is the harmonic mean of the micro-precision and micro-recall

$$Micro - F = \frac{2 * Micro - Pre * Micro - Rec}{Micro - Pre + Micro - Rec} \tag{5.9}$$

Macro-averaged Precision (Santos et al. [2011]) is an average of the precision of the model on different sets

$$Macro - Rec = \frac{\sum_{l=1}^{\mathcal{L}} \frac{TP_l}{TP_l + FP_l}}{\mathcal{L}} \tag{5.10}$$

Macro-averaged Recall (Santos et al. [2011]) is an average of the recall of the model on different sets

$$Macro - Rec = \frac{\sum_{l=1}^{\mathcal{L}} \frac{TP_l}{TP_l + FN_l}}{\mathcal{L}} \tag{5.11}$$

Macro-averaged F-measure (Santos et al. [2011]) is the harmonic mean of the macro-precision and macro-recall

$$Macro - F = \frac{2 * Macro - Pre * Macro - Rec}{Macro - Pre + Macro - Rec} \tag{5.12}$$

5.4.5 Experimental Results

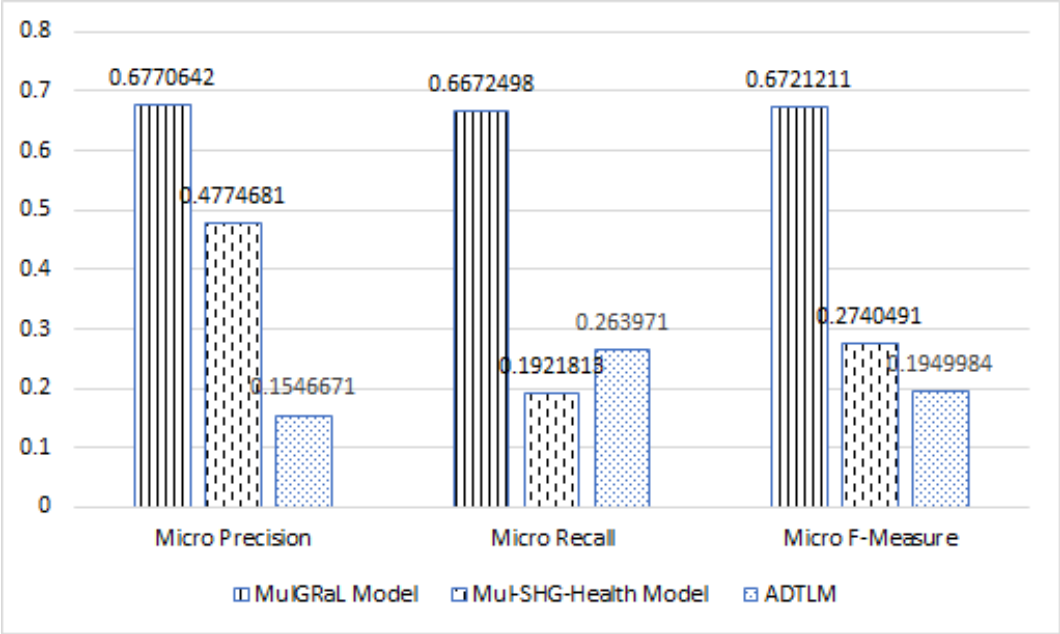


FIGURE 5.7: Micro Comparision between **MulGRaL** and Mul-SHG-Health Model for a large number of diseases

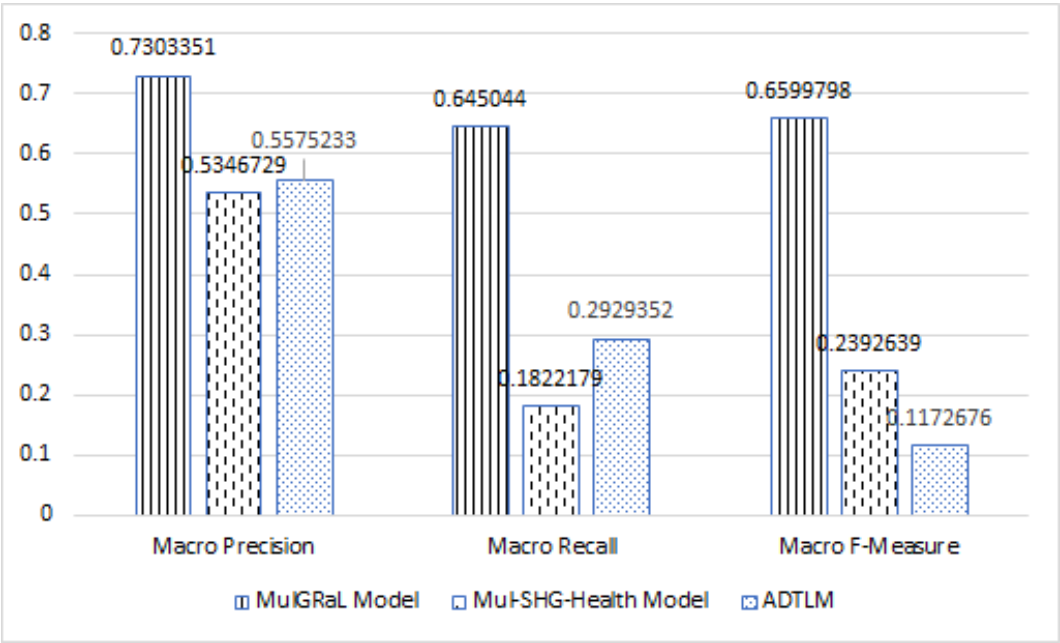


FIGURE 5.8: Macro Comparision between **MulGRaL** and Mul-SHG-Health Model for a large number of diseases

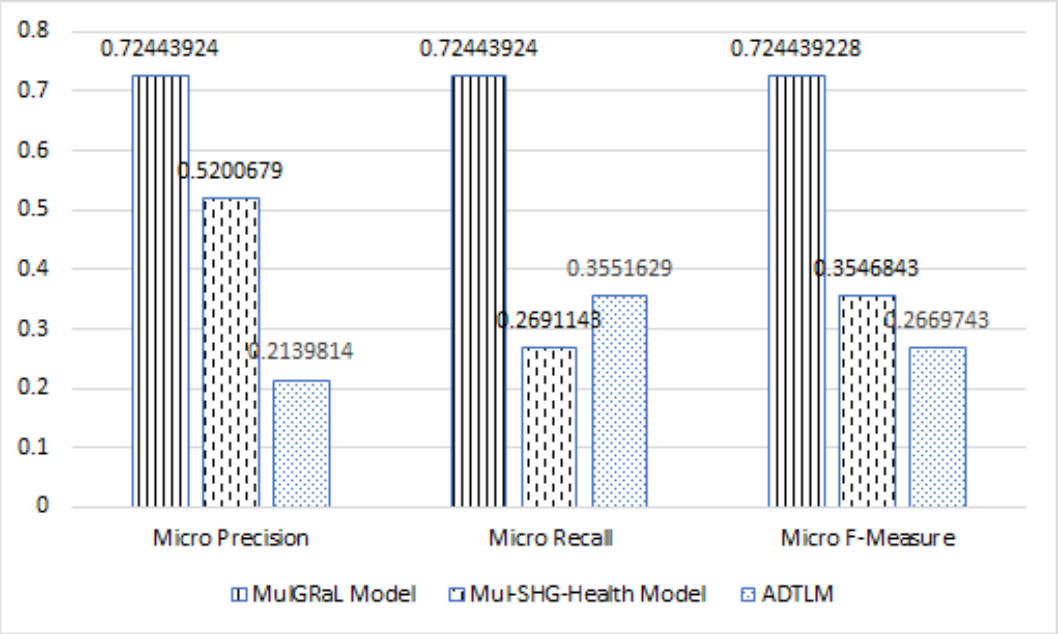


FIGURE 5.9: Micro Comparision between **MulGRaL** and Mul-SHG-Health Model for a medium number of diseases

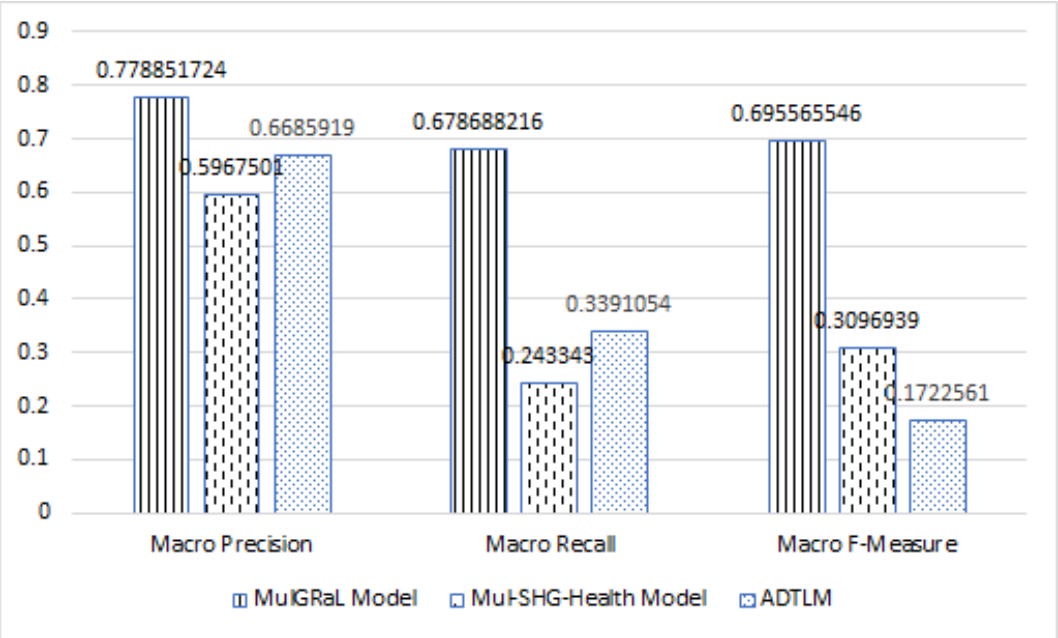


FIGURE 5.10: Macro Comparision between **MulGRaL** and Mul-SHG-Health Model for a medium number of diseases

This study used several performance measures in Section 5.4.4 to evaluate the experimental results. *Figures* 5.7 – 5.12 presented the results of both

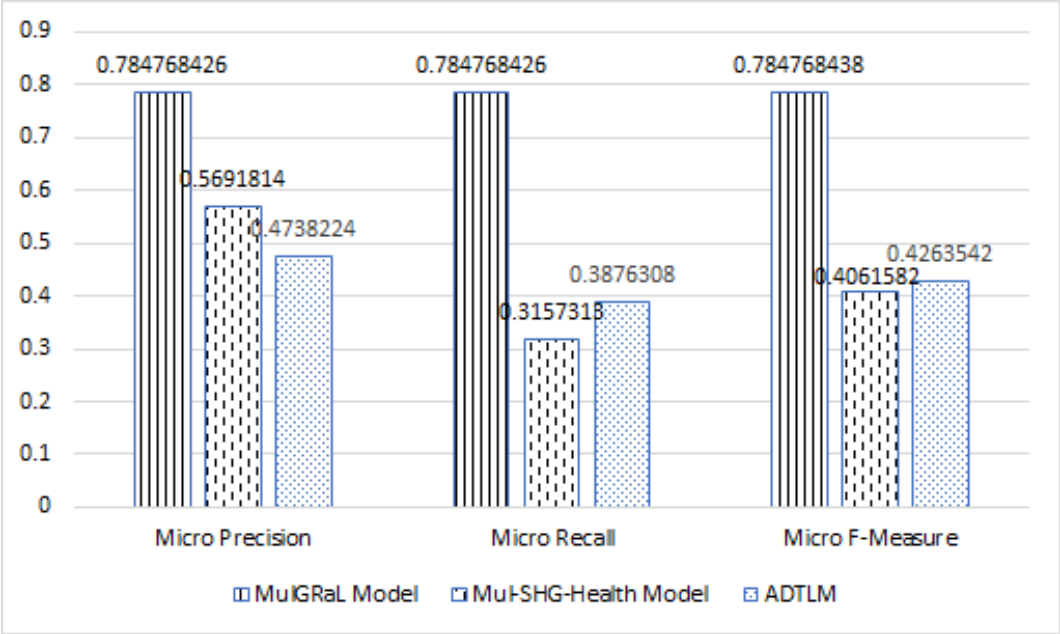


FIGURE 5.11: Micro Comparision between **MulGRaL** and Mul-SHG-Health Model for a small number of diseases

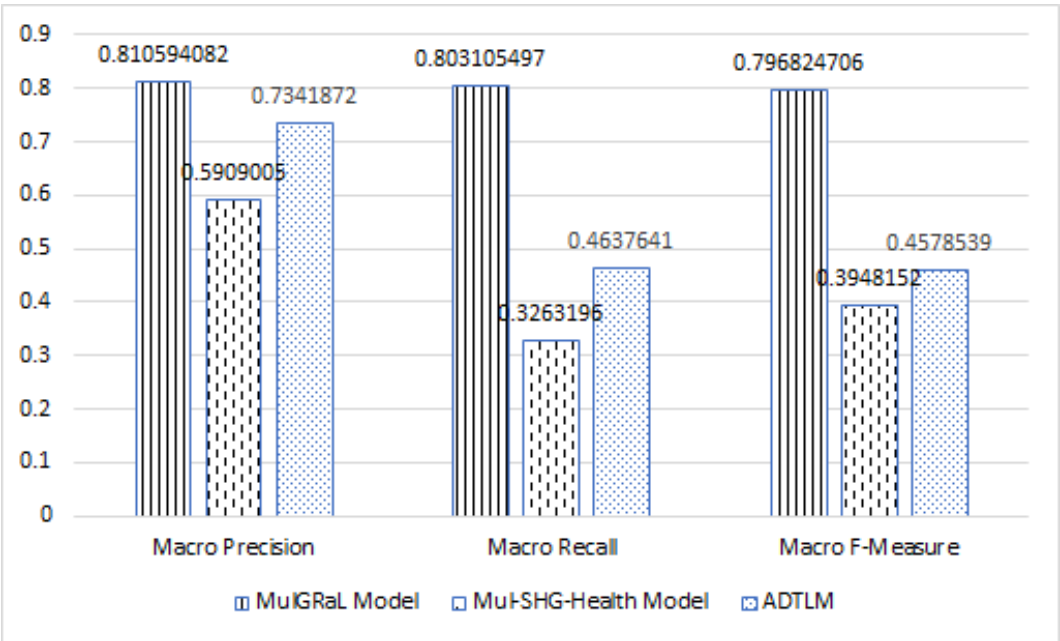


FIGURE 5.12: Macro Comparision between **MulGRaL** and Mul-SHG-Health Model for a small number of diseases

micro and macro regarding precision, recall and F-measure. The figures showed that the performance of classification by MulGRaL was better than

the Mul-SHG-Health and ADTLM model. In particular, the performance of classification through the Macro-F-measure and Micro-F-measure methods for the proposed model established a significant improvement compared to both the Mul-SHG-Health and ADTLM model. 0.6721211 was of Micro-F-measure by the proposed model by a large number of diseases, where 0.2740491 and 0.1949984 was of Micro-F-measure by the Mul-SHG-Health and ADTLM model. The Macro-F-measure for the proposed model, the Mul-SHG-Health model and the ADTLM model was 0.6599798, 0.2392639 and 0.1172676 respectively. There were a significant improvement of performance by a small number of diseases. Both Micro-F-measure and Macro-F-measure of the proposed model achieved 0.7847684 and 0.8105940, respectively. The Mul-SHG-Health model increased to 0.4061582 and 0.3948152, where the ADTLM model raised to 0.4263542 and 0.4578539.

TABLE 5.3: A Comparison of Hamming Loss between **MulGRaL** and Mul-SHG-Health Model for a random set of five subset dataset (5-fold), where the emphasised values indicate the superior performance in comparison.

Hamming Loss	**MulGRaL**	**Mul-SHG-Health**	**Percentage Change**
Round 1	**0.12024**	0.21456	-43.96%
Round 2	**0.11922**	0.21501	-44.55%
Round 3	**0.12016**	0.21293	-43.57%
Round 4	**0.12007**	0.21695	-44.66%
Round 5	**0.11888**	0.21280	-44.13%
Mean	**0.11971**	0.21445	-44.17%

Table 5.3 and 5.4 display the performance of Hamming loss. Hamming loss is a function that shows the rate of the incorrect labelling as compared to the total number of labels; therefore, the best value of hamming loss for a classification model is zero. As mentioned in Section 5.4.1, five subsets

TABLE 5.4: A Comparison of Hamming Loss between **MulGRaL** and ADTLM Model for a random set of five subset dataset (5-fold), where the emphasised values indicate the superior performance in comparison.

Hamming Loss	MulGRaL	ADTLM	Percentage Change
Round 1	**0.12024**	0.17044	-29.45%
Round 2	**0.11922**	0.16228	-26.53%
Round 3	**0.12016**	0.17715	-32.17%
Round 4	**0.12007**	0.16613	-27.73%
Round 5	**0.11888**	0.16817	-29.31%
Mean	**0.11971**	0.16884	-29.09%

of the dataset were randomly selected to evaluate the percentage of the incorrect and correct labels. An average of five rounds was the final result of the experiment. *Table* 5.3 and 5.4 show that the prediction of failures in this study was 0.11971, where the Mul-SHG-Health model was 0.21445 and the ADTLM model was 0.16884, which indicates that the performance of **MulGRaL** is better than the Mul-SHG-Health model. The result shows that the rate of the incorrect predictions in **MulGRaL** was reduced by -44.17% compared to the Mul-SHG-Health model and -29.09% compared to the ADTLM model. The measure for identifying the percentage change in performance was based on the *Equation* 3.10.

Table 5.5 and 5.6 display the performance of Subset Accuracy. Subset Accuracy shows the correct label of prediction for the sample label in both possible and negative classifiers and the optimal value is one. The correct label of classification compared to all samples for the prosed model was also higher than the Mul-SHG-Health model. *Table* 5.5 illustrates that **MulGRaL** improved by 36.85% in comparison to the Mul-SHG-Health model. In contrast, *Table* 5.6 showed that **MulGRaL** has not improved in comparison to

TABLE 5.5: A Comparison of subset accuracy between **MulGRaL** and Mul-SHG-Health Model for a random set of five subset dataset (5-fold), where the emphasised values indicate the superior performance in comparison.

Subset Accuracy	**MulGRaL**	**Mul-SHG-Health**	**Percentage Change**
Round 1	**0.49365**	0.35711	38.24%
Round 2	**0.48550**	0.35892	35.27%
Round 3	**0.49094**	0.35484	38.35%
Round 4	**0.48528**	0.35529	36.58%
Round 5	**0.48867**	0.35982	35.81%
Mean	**0.48881**	0.35720	36.85%

TABLE 5.6: A Comparison of subset accuracy between **MulGRaL** and ADTLM Model for a random set of five subset dataset (5-fold), where the emphasised values indicate the superior performance in comparison.

Subset Accuracy	**MulGRaL**	**ADTLM**	**Percentage Change**
Round 1	**0.49365**	0.48460	1.87%
Round 2	**0.48550**	0.50090	-3.07%
Round 3	**0.49094**	0.47826	2.65%
Round 4	**0.48528**	0.50634	-4.16%
Round 5	**0.48867**	0.49184	-0.64%
Mean	**0.48881**	0.49239	-0.67%

the ADTLM model, however, this is just a small percent with only -0.67%.

5.4.6 Discussions

Impact of Using Knowledge Graph in Multi-label Classification

Learning multi-label classification is a complex and challenging research task. This study exploits multi-label problems by applying a heterogeneous information graph. The graph is populated with the knowledge learned

from the medical domain. Integrating the knowledge graph into the classification model yields a significant effect on the performance. This advantage was demonstrated in Section 5.4.5. The advantage of a knowledge graph was its capability to discover hidden relationships among objects, which helped to improve learning multi-labels by label correlation. The approach increased the ability to identify the relationship among labels based on semantic relations. Experimental results of the proposed model were better than the Mul-SHG-Health model, which showed that the knowledge base played a vital role in improving the accuracy of prediction for learning classification.

Impact of Level Disease in the Multi-label Classification

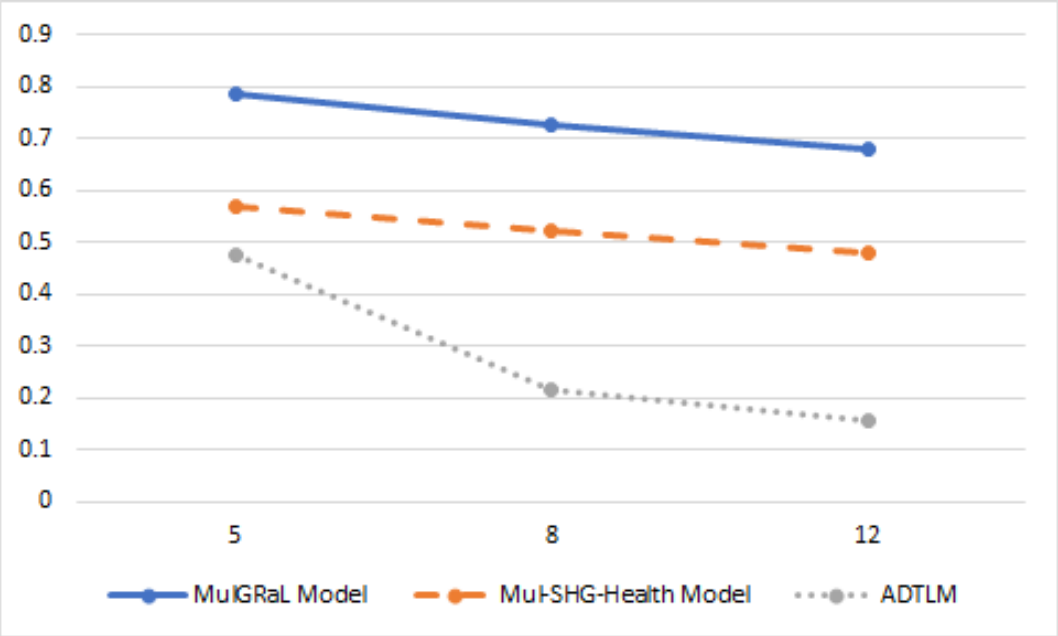


FIGURE 5.13: Micro-Precision Comparison between **MulGRaL**, ADTLM and Mul-SHG-Health model with different number of diseases

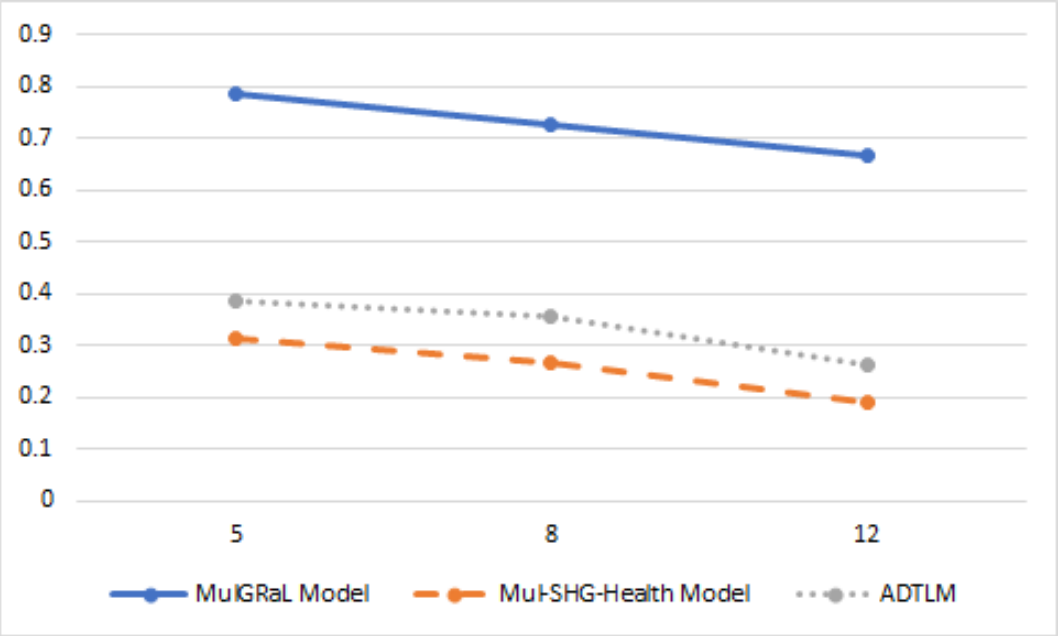


FIGURE 5.14: Micro-Recall Comparison between **MulGRaL**, ADTLM and Mul-SHG-Health model with different number of diseases

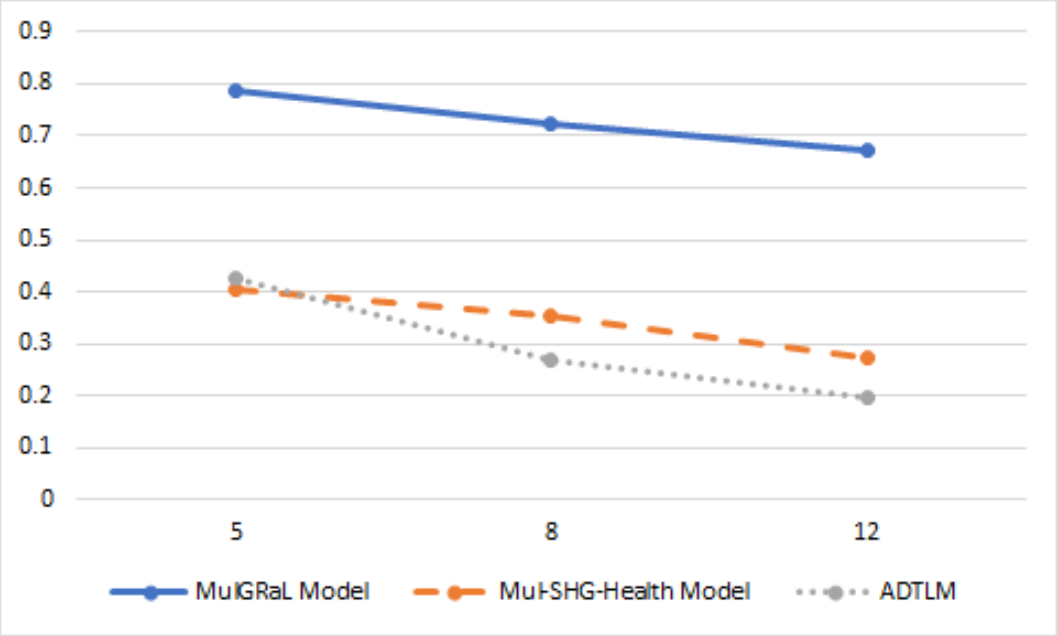


FIGURE 5.15: Micro-FMeasure Comparison between **MulGRaL**, ADTLM and Mul-SHG-Health model with different number of diseases

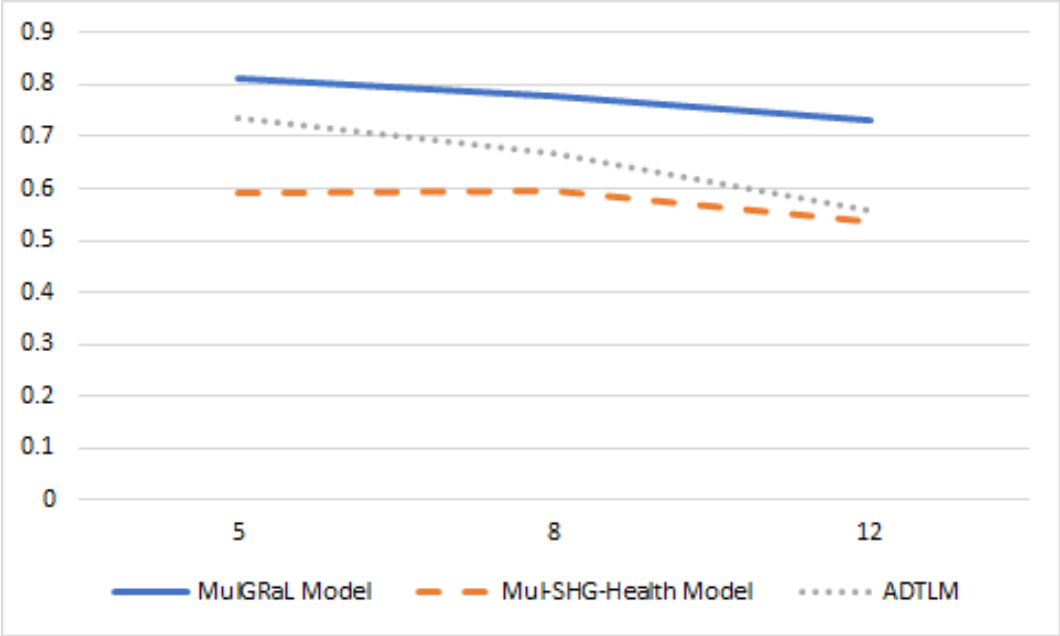


FIGURE 5.16: Macro-Precision Comparison between **MulGRaL**, ADTLM and Mul-SHG-Health model with different number of diseases

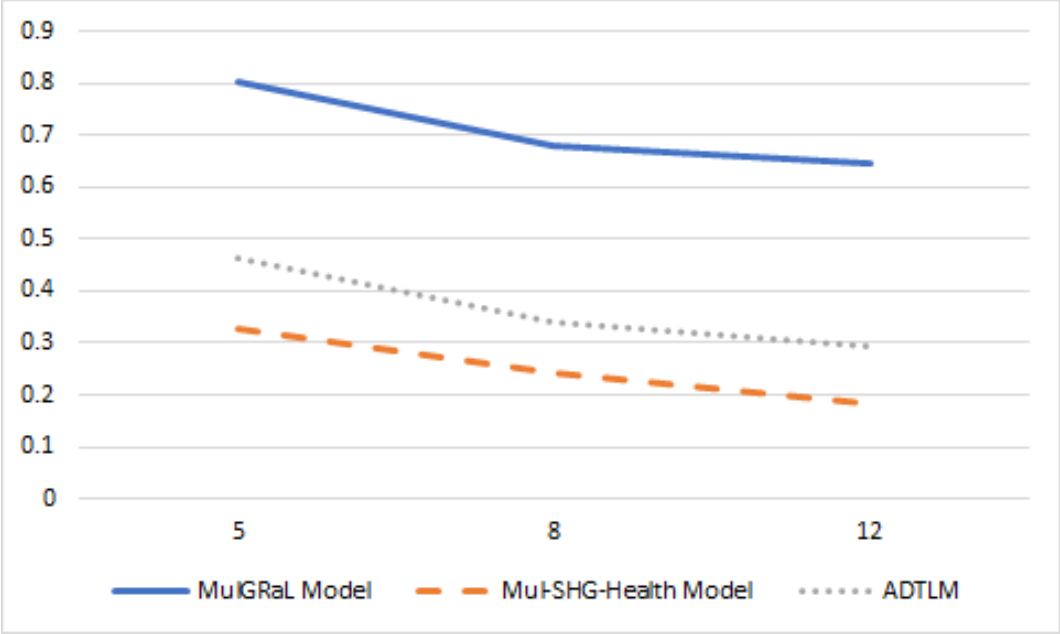


FIGURE 5.17: Macro-Recall Comparison between **MulGRaL**, ADTLM and Mul-SHG-Health model with different number of diseases

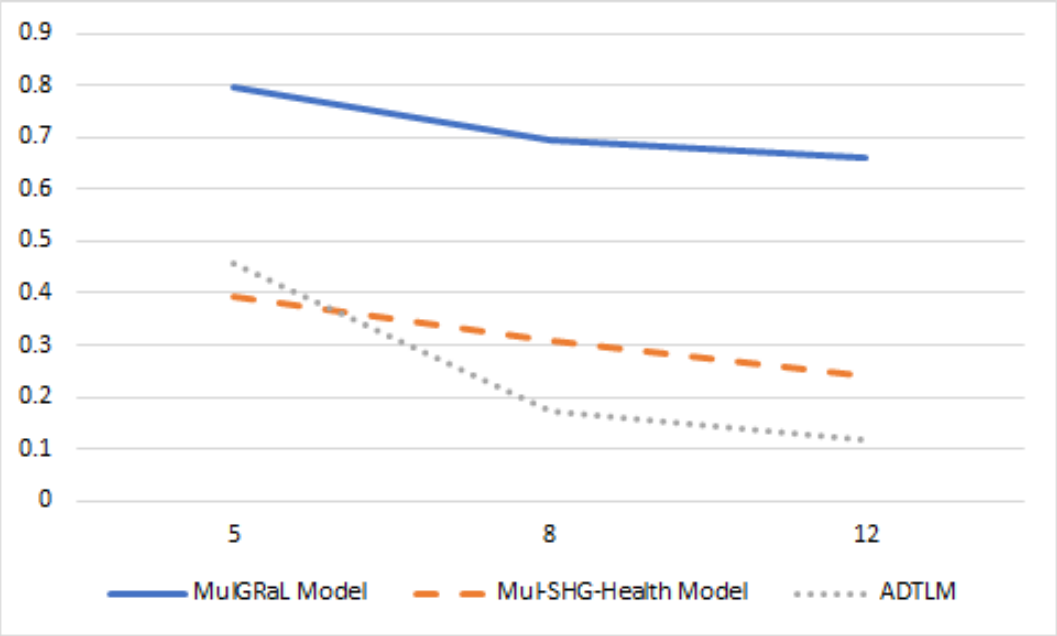


FIGURE 5.18: Macro-FMeasure Comparison between **MulGRaL**, ADTLM and Mul-SHG-Health model with different number of diseases

As mentioned in Section 5.4.2, a three-set layered hierarchy of the number of diseases, including five diseases, eight diseases and twelve diseases was used to test the performance of classification. Each layer corresponded to each round of the experiment. Round 1 used the most popular five of twelve diseases (MCQ160A, MCQ010, DIQ010, MCQ160M,MCQ160K) in Figure 5.4. Round 2 was run with eight diseases: MCQ160A, MCQ010, DIQ010, MCQ160M, MCQ160K, MCQ160N, MCQ160L, MCQ160C. Round 3 used twelve diseases for the experiment. The result of this proposed approach shows that when more disease sets were used for learning classification; the performance of classification decreased. Figure 5.13, 5.14, 5.15, 5.16, 5.17 and 5.18 present the performance of the classification model regarding

micro and macro levels for three-level disease sets. The performance of classification for twelve diseases decreased compared to five diseases. This result may be because the label of the new label set increased, which led to the inability to distinguish among the computational label set.

TABLE 5.7: A Comparison of Hamming Loss between **MulGRaL** and Mul-SHG-Health Model for three subset disease, where the emphasised values indicate the superior performance in comparison.

Hamming Loss	**MulGRaL**	**Mul-SHG-Health Model**	**Percentage Change**
Round 1	**0.11586**	0.24395	-52.50%
Round 2	**0.13113**	0.21975	-40.33%
Round 3	**0.11971**	0.21445	-44.17%

TABLE 5.8: A Comparison of Hamming Loss between **MulGRaL** and ADTLM Model for three subset disease, where the emphasised values indicate the superior performance in comparison.

Hamming Loss	**MulGRaL**	**ADTLM**	**Percentage Change**
Round 1	**0.11586**	0.19188	-39.61%
Round 2	**0.13113**	0.18926	-30.72%
Round 3	**0.11971**	0.16884	-29.09%

Table 5.7 and 5.8 demonstrates that the percentage of prediction failures of high-level diseases was higher than the low-level diseases. Similarly, Table 5.9 and 5.10 shows various effects on the different levels of disease sets. When the first five diseases, as in Figure 5.4, were used for experimentation, attributes of the dataset were apparent, which generated a higher quality training dataset with lesser noise. Therefore, since a higher quality training dataset was used for the experimentation, the predicted ratio of correct

TABLE 5.9: A Comparison of subset accuracy between **MulGRaL** and Mul-SHG-Health Model for three subset disease, where the emphasised values indicate the superior performance in comparison.

Subset Accuracy	**MulGRaL**	**Mul-SHG-Health Model**	**Percentage Change**
Round 1	**0.72287**	0.40217	79.74%
Round 2	**0.52350**	0.38378	36.41%
Round 3	**0.48881**	0.35720	36.85%

TABLE 5.10: A Comparison of subset accuracy between **MulGRaL** and ADTLM Model for three subset disease, where the emphasised values indicate the superior performance in comparison.

Subset Accuracy	**MulGRaL**	**ADTLM**	**Percentage Change**
Round 1	**0.72287**	0.59311	21.88%
Round 2	**0.52350**	0.52590	-0.46%
Round 3	**0.48881**	0.49239	-0.73%

labels was higher. When the highest level of disease was used, the related relationships were of lower quality and contained more noise. With more attributes added into the class set, the whole experimental dataset was noisier and required more computational resources of classifying for the classification model. Therefore, the performance of classification decreased because both the number of the valid samples and the increase of ambiguous information was small.

Optimization

In this study, an Algorithm 5.2 in Section 5.3.4 was proposed to determine the learning of multi-label classification. Computing and processing for a

dataset with a large number of labels take a lot of time and cost because of the increase in combinational label space. The drawback degraded the model's ability to classify because of ambiguity among various labels. Therefore, optimisation for the algorithm was requested to help improve the performance of the classification model. This study introduced a threshold of α for optimising the result of the prediction. This coefficient played a vital role in determining possible labels. The performance of predicting correct labels for its classifier decreased if there was multiple similar options for the prediction. The model randomly selected any label if it could not distinguish among labels. Therefore, the threshold α aimed to discriminate between labels which helped advance the capability of classification. This value depends the size of the dataset when used for our training model. The value of α is in the range zero to λ, where λ is calculated as the average of the residual of all combinational labels $\acute{\mathcal{L}}$. There is a loop running from zero to λ. The value for each run is 0.01. Table 5.11 shows an example of the experiment by a small number of diseases with different values of α. Based on this experiment, the best performance of the classification model is identified by the threshold of α equalling to 0.02.

TABLE 5.11: Sample comparison of different values of α in Experiments.

(α)	Hamming-Loss	Subset Accuracy	Macro-Precision	Macro-Recall	Macro-Fmeasure
0.01	0.11793	0.71739	0.80312	0.79619	78774
0.02	0.11512	0.72305	0.81027	0.80270	79597
0.03	0.12210	0.70969	0.79517	0.78726	77660
0.04	0.12699	0.69927	0.78357	0.78068	76275
0.05	0.12862	0.69565	0.78050	0.78836	75915

5.5 Summary

Exploiting the relevance label plays a vital role in boosting the performance of multi-label classification. Therefore, a large number of researchers have focused on this approach to deal with the multi-label classification. However, this method increases the label space set during learning. This study presents a multi-label graph method to discover cooperation among labels as well as reduce the space of the label set. The graph can be used to exploit more hidden relationships among labels, which facilitates the discovery of label relevance. In addition, this study enriches knowledge of the graph from other sources to help eliminate inappropriate relationships. Furthermore, separating the graph into a two space network helps the classification model to facilitate the learning label relevance. Finally, a ranking algorithm was proposed to deal with the multi-label classification problem. Experimental results show that the performance of the classification for the proposed model was better than the Mul-SHG-Health and ADTLM model.

Chapter 6

Conclusions

This chapter will present the conclusions of the whole thesis including contributions and significance. Finally, this chapter will discuss a number of points opened for future work.

6.1 Conclusions

In this thesis, the research has focussed on dealing with the problem of classification. The challenges have been divided into three subproblem throughout the thesis. Firstly, tackling with the binary classification, the study used a heterogeneous information graph to discover knowledge from the observational data. The study processed and mined data sources from electronic health records to build a heterogenous information graph. The challenge of this study is how to optimise the exploration of the observation data as well as analyse and evaluate feature factors for developing a heterogenous information graph to support health risk prediction. This problem has been addressed in Chapter 3. In this Chapter, a classification model for risk assessment on healthcare data was developed based on the heterogeneous information graph. Mining on the heterogeneous graph helped to discover

more underlying patterns from the observational data. The National Health and Nutrition Examination Survey dataset was also used and divided into eight semantic groups to facilitate knowledge discovery. The *semantic similarity* was used to ascertain the relations among categories. The method helped to discover more relationships, which provided positive effects on health risk prediction. In addition, the study used the *Pearson Correlation* coefficient to set up the links for building the graph. A classification model later based on the heterogeneous information graph was developed to predict a patient's disease status. The experimental results showed that the proposed model performed better than the baseline model.

Secondly, the study focused on dealing with the problem of integrating medical knowledge into the heterogenous information graph to improve the performance of classification models; this is present in Chapter 4. A knowledge graph was built based on Medical Subject Headings (MeSH) and MEDLINE. MeSH includes a list of concepts that link to the documents of MEDLINE. Each concept is associated with several articles in the medical domain. MEDLINE uses MeSH for indexing of articles. MEDLINE holds approximately 27 million articles in the fields of biomedicine and health and updates frequently. Mining data from MEDLINE helped to discover more relationships, which were hidden between entities, to deal with ambiguous meaning. In this chapter, the weight of the link for the heterogeneous information graph was normalised by instances learned from MEDLINE. There was no direct link between MEDLINE and NHANES. Therefore, this study used the Medical Subject Heading and the International Classification of Diseases as a bridge to map MEDLINE and NHANES. To connect MEDLINE and NHANES for populating the knowledge graph, a mapping

function of MEDLINE, MeSH and NHANES was used. After obtaining the mapping, the research extracted titles and abstracts related to a specific topic. The extracted data was used to identify the weight of terms related to that topic by the vector space model which was used to generate feature vectors from documents in MEDLINE. These terms were later linked with nodes from the heterogeneous information graph as instances to normalise the edge of the graph. After populating the knowledge for the heterogeneous information graph by using instances, a classification model was constructed to mine the knowledge graph for predicting the health risk status. The result of this proposal model brought a significant improvement, which contributed to an increase in the accuracy of the prediction. This study also demonstrated that applying the knowledge base into the classification model has achieved more benefits than others without using the knowledge base.

Finally, the study addressed the problem of multi-label prediction in chapter 5. The researcher proposed a framework based on the observation data in healthcare to discover the status of patients with multi-risks. A heterogeneous information graph was adopted with the observation data to exploit the correlation among labels — each label class was associated with a subgraph with extracts from the heterogeneous information graph. The subgraph was normalised by medical knowledge that was learned from MEDLINE. Mapping between observation data and MEDLINE helped to exploit more hidden information to learn label correlation. Exploiting the relevance label played a vital role in boosting the performance of multi-label classification. With the learning classification established on the knowledge base, the imbalance data and missing data could be reduced. Especially in

the space of the heterogeneous information graph, discovering cooperation among different labels and linkages helped to exploit more hidden correlational among labels. The approach had a significant effect on developing the multi-label classification model (Each patient may have several diseases, and these diseases may have some similar or different symptoms. If a classifier is not capable of linking all attributes of patients together, the result of predictions may not achieve high performance). Furthermore, separating the graph into two space samples was used to improve the effect of learning label relevance. An algorithm of the combined label set was later proposed to learn correlation among labels as a new label set. Finally, a rank label algorithm was introduced to predict the new combinational label set correspondent as multiple label graph. Experimental results showed that the performance of the classification for the proposed model was better than the baseline model.

6.2 Scientific Contributions and Significance

Dealing with the classification problem can bring advantages of developing decision-support systems. The thesis suggested several innovative methods to solve the problem of learning classification. The research mined the observational data and medical domain corpus in healthcare to develop classification models, which aimed to support practitioners in reducing human errors. The key operation of the method was to use a heterogeneous information graph to represent the data for achieving high performance of discovering knowledge. The graph also integrated a medical knowledge base that mined from the citations of published papers in MEDLINE to discover

more high relationships among objects in the observational data. These approaches brought advantages in developing classification models with high performance.

The main novelty of this thesis is a combination of *Pearson correlation* and *semantic classes* to build a heterogeneous information graph. This graph later has populated knowledge in medical domain that learned from MEDLINE. MEDLINE is a bibliographic database that has collected journal articles from academic journals in life sciences and biomedical information since 1966. It is produced by the National Library of Medicine in the United States. These academic journals cover medicine, veterinary medicine, nursing, dentistry, pharmacy, and health care. The database contains more than 27 million references which are selected from more than 5200 international publications in about 40 languages (Costa et al. [2018]). Based on this graph, the binary classification model, as well as multi-label classification model, is developed to predict the health risk status.

This study also has obtained a high-level for the reliability of the system. Firstly, this study has used two real datasets (NHANES and NAMCS) to develop classification models. Besides, the model has used a bibliographic database (MEDLINE) as evidence-based to improve the accuracy of the classification model. In addition, this study has used two baseline model (Chen et al. [2016a], Wang et al. [2020]) which known as the state-of-the-art of the technology for health risk prediction.

The thesis made significant contributions in the following:

- A new binary classification model was proposed, which provided a solution as a classification model to assess personal health status (Pham et al. [2018]). The classification model considers using the *Semantic*

similarity to health risk predictions by discovering knowledge from health examination records. The model was developed based on a heterogeneous information graph which was built by using the *Pearson Correlation* coefficient for setting up the link between two nodes. Presenting the framework to create a heterogeneous information graph also played an essential role in developing a classification model to assess personal health status.

- The thesis introduced a framework to build a knowledge base (Pham et al. [2019]) that could help to manage more meaningful information — providing a knowledge base in the medical domain that could improve the result of searching based on semantic relationships among entities. The study later applied the medical knowledge graph in a classification model to improve the performance of the classification model(Pham et al. [2020]). Using the knowledge graph helped researchers to discover new medical knowledge and achieve a deep understanding of subjects in the medical domain. Therefore, applying a knowledge base into the classification model helped improve the quality of prediction of the health risk status.

- A framework of the multi-label graph was proposed for dealing with the multi-label classification problem. This method also considered separating the graph into positive and negative spaces. The approach helped to improve the discovery of the relevance label. Later, a ranking algorithm based on this framework was introduced for predicting multiple diseases. The experimental result for both Hamming-loss and Subset-Accuracy achieved a significant improvement compared to the state-of- the-art model with 44,17% and 36,85% as in Table 5.3

and 5.5, respectively. The study helped to decrease challenges in learning the multi-label classification model, a research area that has increased quickly in real-life.

- The thesis contributed to the social effect in the medical domain, which helped reduce financial and timing costs for physicians and healthcare practitioners. The practical use of data mining techniques in supporting the diagnosis of disease has the potential to improve the quality of healthcare. The practicality of the system is its ability to give practitioners more professional advice, thus helping them to reduce the potential risk of human errors.

- The thesis discovered new medical knowledge by combining of observation data, MeSH and MEDLINE. Rebuilding the structures of MeSH and MEDLINE for building a knowledge graph contributed a useful data source for researchers and developers in developing applications related to a treatment plan or decision support.

- The thesis made significant contributions to the advancement of knowledge in data mining with an innovative classification model specifically crafted for domain-based data. Moreover, the research contributed to the healthcare community by providing a deep understanding of people's health with accessibility to the patterns in various observations.

6.3 Future Work

In the study, using the heterogeneous information graph helped to deal with the imbalance of data and exploit the ambiguous meaning of entities in the

real dataset. Moreover, applying *Pearson correlation* in discovery links of entities contributed to advance the performance of the classification model. The proposed model within the knowledge base helped to increase the accuracy as well as the reliability of the classification model. Knowledge discovery from science literature has played an essential role in healthcare. As a result, these models have capabilities in supporting doctors to avoid human errors, which achieved significant contributions to improve the quality of healthcare. For future work, expanding the present investigation is expected to explore the following aspects.

The first issue is the time series data of Health Records. For diagnosis, time series plays a vital role in determining the outcome of the prediction. Each patient may have characteristics related to each type of disease from time to time. As a result, the relationship between features and conditions may change. In the diagnosis of the disease, the pair of cause and effect relationships are essential and will impact both directly and indirectly on conditions from time to time. However, converting and manipulating data types over time for classification models is still an open challenge. Therefore it is important to study the data type over-time of a dataset to help improve the outcome of the prediction.

Secondly, in the field of treatment and diagnosis, updated data is essential. Using the dataset without an update to evaluate the health risk status may have a negative effect on the current predictions. Therefore, the heterogeneous information should be updated. Linking existing data with social data network may have a positive effect on the performance of the classification model. In particular, linking the classification model to experienced

doctors to justify the results may help improve the accuracy of the classification model. More evaluation from experts will lead to more accurate and useful classification model. It is expected that this approach will increase in popularity, having a positive impact on classification models in general as well as on diagnosis of particular disease. However, incorporating social networking data into classification models as well as an expert intervention into the system to evaluate predictive results will not be an easy task for researchers. Therefor, future researchers will need to pay more attention to this challenge to yield an effective predictive result. The heterogeneous information graph is also expected to support the development of knowledge engineering and knowledge management.

Thirdly, discovery of the relevance among features and labels is expected increase the accuracy of the work that is done through feature extraction and feature selection. This approach will help to provide a more solid basis for the current model. Deep learning could be an effective approach to deal with this issue. Deep learning is one of the Machine Learning techniques that are based on learning multiple levels of abstraction and representation. Deep learning techniques comprise multiple processing layers within a computational model to learn representations of data with various levels of abstraction. Deep learning uses more hidden layers than traditional machine learning to draw out representations from the raw data. Miotto et al. [2017] argued that diverse datasets could be integrated across heterogeneous data types by using deep architectures. Such relationships can be generated for a range of semantic relations as well as a syntactic relation. Therefore, learning presentation data using deep learning has the ability to understand the meaning hidden inside data, which can help to learn label

relevance.

Bibliography

Lisa B Mirel and Kelly Carper. Trends in health care expenditures for the elderly, age 65 and older: 2001, 2006, and 2011. In *Statistical Brief (Medical Expenditure Panel Survey (US))[Internet]*. Agency for Healthcare Research and Quality (US), 2014. URL `https://www.ncbi.nlm.nih.gov/books/NBK476262/`.

Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

Maria C Inacio, Sarah Catherine Elizabeth Bray, Craig Whitehead, Megan Corlis, Renuka Visvanathan, Keith Evans, Elizabeth C Griffith, and Steve L Wesselingh. Registry of older south australians (rosa): framework and plan. *BMJ open*, 9(6):e026319, 2019.

Hian Chye Koh, Gerald Tan, et al. Data mining applications in healthcare. *Journal of healthcare information management*, 19(2):65, 2011.

Paul A James, Suzanne Oparil, Barry L Carter, William C Cushman, Cheryl Dennison-Himmelfarb, Joel Handler, Daniel T Lackland, Michael L LeFevre, Thomas D MacKenzie, Olugbenga Ogedegbe, 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). *The Journal of the American Medical Association (JAMA)*, 311(5):507–520, 2014.

MG Myriam Hunink, Milton C Weinstein, Eve Wittenberg, Michael F Drummond, Joseph S Pliskin, John B Wong, and Paul P Glasziou. *Decision making in health and medicine: integrating evidence and values*. Cambridge University Press, 2014.

Cheng-Ding Chang, Chien-Chih Wang, and Bernard C Jiang. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. *Expert systems with applications*, 38(5):5507–5513, 2011.

Feixiang Huang, Shengyong Wang, and Chien-Chung Chan. Predicting disease by using data mining based on healthcare information system. In *2012 IEEE International Conference on Granular Computing*, pages 191–194. IEEE, 2012.

Jae-Kwon Kim, Jong-Sik Lee, Dong-Kyun Park, Yong-Soo Lim, Young-Ho Lee, and Eun-Young Jung. Adaptive mining prediction model for content recommendation to coronary heart disease patients. *Cluster computing*, 17 (3):881–891, 2014.

M Sabibullah, V Shanmugasundaram, and R Priya. Diabetes patient's risk through soft computing model. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2(6):60–65, 2013.

Jia Zhang, Candong Li, Donglin Cao, Yaojin Lin, Songzhi Su, Liang Dai, and Shaozi Li. Multi-label learning with label-specific features by resolving label correlations. *Knowledge-Based Systems*, 159:148–157, 2018a.

Setu Shah, Xiao Luo, Saravanan Kanakasabai, Ricardo Tuason, and Gregory Klopper. Neural networks for mining the associations between diseases and symptoms in clinical notes. *Health information science and systems*, 7 (1):1, 2019.

Huirui Han, Mengxing Huang, Yu Zhang, Xiaogang Yang, and Wenlong Feng. Multi-label learning with label specific features using correlation information. *IEEE Access*, 7:11474–11484, 2019.

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*, 28 (9):2423–2437, 2016a.

Yun Xiong, Lu Ruan, Mengjie Guo, Chunlei Tang, Xiangnan Kong, Yangyong Zhu, and Wei Wang. Predicting disease-related associations by heterogeneous network embedding. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 548–555. IEEE, 2018.

Xiujuan Lei and Yuchen Zhang. Predicting disease-genes based on network information loss and protein complexes in heterogeneous network. *Information Sciences*, 479:386–400, 2019.

Tingyan Wang, Robin G Qiu, Ming Yu, and Runtong Zhang. Directed disease networks to facilitate multiple-disease risk assessment modeling. *Decision Support Systems*, 129:113171, 2020.

Haolin Wang, Qingpeng Zhang, and Jiahu Yuan. Semantically enhanced medical information retrieval system: a tensor factorization based approach. *IEEE Access*, 5:7584–7593, 2017a.

Ming Ji, Jiawei Han, and Marina 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. ACM, 2011a.

Yizhou Sun, Yintao Yu, and Jiawei 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.

J David. Principles of data mining. *Drug safety*, 30(7):621–622, 2007.

Matthew Herland, Taghi M Khoshgoftaar, and Randall Wald. A review of data mining using big data in health informatics. *Journal of Big data*, 1(1): 1–35, 2014.

Saharon Rosset, Claudia Perlich, Grzergorz Świrszcz, Prem Melville, and Yan Liu. Medical data mining: insights from winning two competitions. *Data Mining and Knowledge Discovery*, 20(3):439–468, 2010.

Andreas Holzinger. Machine learning for health informatics. In *Machine Learning for Health Informatics*, pages 1–24. Springer, 2016.

Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4): 2431–2448, 2012.

Y Miyazaki. Revised international prognostic scoring system (ipss-r) for myelodysplastic syndromes. *[Rinsho ketsueki] The Japanese journal of clinical hematology*, 54(6):545, 2013.

Prashant Prakash, Kavita Krishna, and Deepansh Bhatia. Usefulness of saps ii scoring system as an early predictor of outcome in icu patients. *Journal Indian Academy of Clinical Medicine (JIACM)*, 7(3):202–5, 2006.

Douglas P Wagner and Elizabeth A Draper. Acute physiology and chronic health evaluation (apache ii) and medicare reimbursement. *Health care financing review*, 1984(Suppl):91, 1984.

Mark T Keegan, Ognjen Gajic, and Bekele 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.

Duen-Yian Yeh, Ching-Hsue Cheng, and Yen-Wen Chen. A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*, 38(7):8970–8977, 2011.

Hani Neuvirth, Michal Ozery-Flato, Jianying Hu, Jonathan Laserson, Martin S Kohn, Shahram Ebadollahi, and Michal Rosen-Zvi. Toward personalized care management of patients at risk: the diabetes case study. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 395–403. ACM, 2011.

Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification models with soft-label information. *Journal of the American Medical Informatics Association*, 21(3):501–508, 2014.

Yazhou Yang and Marco Loog. Active learning using uncertainty information. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2646–2651. IEEE, 2016.

Andrew Guillory and Jeff A Bilmes. Label selection on graphs. In *Advances in Neural Information Processing Systems*, pages 691–699, 2009.

Ming Ji, Yizhou Sun, Marina Danilevsky, Jiawei Han, and Jing Gao. Graph regularized transductive classification on heterogeneous information networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Springer, 2010.

Chen Luo, Renchu Guan, Zhe Wang, and Chenghua Lin. Hetpathmine: A novel transductive classification algorithm on heterogeneous information networks. In *European Conference on Information Retrieval*, pages 210–221. Springer, 2014.

Taehyun Hwang and Rui Kuang. A heterogeneous label propagation algorithm for disease gene discovery. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 583–594. SIAM, 2010.

Jin-Bok Lee, Jung-jae Kim, and Jong C Park. Automatic extension of gene ontology with flexible identification of candidate terms, 2006.

Wei Gao, Mohammad Reza Farahani, Adnan Aslam, and Sunilkumar Hosamani. Distance learning techniques for ontology similarity measuring and ontology mapping. *Cluster Computing*, 20(2):959–968, 2017.

Yuan Ni, Qiong Kai Xu, Feng Cao, Yosi Mass, Dafna Sheinwald, Hui Jia Zhu, and Shao Sheng Cao. Semantic documents relatedness using concept graph representation. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 635–644. ACM, 2016.

Sungbin Choi, Jinwook Choi, Sooyoung Yoo, Heechun Kim, and Youngho Lee. Semantic concept-enriched dependence model for medical information retrieval. *Journal of biomedical informatics*, 47:18–27, 2014.

Chunye Wang and Ramakrishna Akella. Concept-based relevance models for medical and semantic information retrieval. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 173–182. ACM, 2015.

P Karpagam, S Sivasubramanian, and C Nalini. Extending disease ontology with newly evaluated terms to improve semantic medical information retrieval. *International Journal of Applied Engineering Research*, 11(5): 3527–3535, 2016.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1219–1228. ACM, 2014a.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 301–306, 2011.

Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, and Nathalie Souf. Learning concept-driven document embeddings for medical information search. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 160–170. Springer, 2017.

Longxiang Shi, Shijian Li, Xiaoran Yang, Jiaheng Qi, Gang Pan, and Binbin Zhou. Semantic health knowledge graph: Semantic integration of heterogeneous medical knowledge and services. *BioMed research international*, 2017. doi: https://doi.org/10.1155/2017/2858423.

Nikos Voskarides, Edgar Meij, Manos Tsagkias, Maarten De Rijke, and Wouter Weerkamp. Learning to explain entity relationships in knowledge graphs. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 564–574, 2015.

Suzanne Pereira, Aurélie Névéol, Philippe Massari, Michel Joubert, and Stefan Darmoni. Construction of a semi-automated icd-10 coding help system to optimize medical and economic coding. In *Medical Informatics Europe (MIE)*, pages 845–850, 2006.

Suresh Srinivasan, Thomas C Rindflesch, William T Hole, Alan R Aronson, and James G Mork. Finding umls metathesaurus concepts in medline. In *Proceedings of the AMIA Symposium*, pages 727–731. American Medical Informatics Association, 2002.

Lynn M Schriml, Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, Linda Jeng, Cynthia Bearer, Richard Lichenstein, et al. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1):D955–D962, 2018.

Himali Saitwal, David Qing, Stephen Jones, Elmer V Bernstam, Christopher G Chute, and Todd R Johnson. Cross-terminology mapping challenges: a demonstration using medication terminological systems. *Journal of biomedical informatics*, 45(4):613–625, 2012.

Yuanyuan Zhang, Pradip K Srimani, and James Z Wang. Combining mesh thesaurus with umls in pseudo relevance feedback to improve biomedical information retrieval. In *2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA)*, pages 67–71. IEEE, 2016.

Huda Banuqitah, Fathy Eassa, Kamal Jambi, and Maysoon Abulkhair. Two level self-supervised relation extraction from medline using umls. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 6 (3):11–23, 2016.

Yicheng Jiang, Bensheng Qiu, Chunsheng Xu, and Chuanfu Li. The research of clinical decision support system based on three-layer knowledge base model. *Journal of healthcare engineering*, 2017(6535286):1–8, 2017.

Liqin Wang, Guilherme Del Fiol, Bruce E Bray, and Peter J Haug. Generating disease-pertinent treatment vocabularies from medline citations. *Journal of biomedical informatics*, 65:46–57, 2017b.

Zhisheng Huang, Jie Yang, Frank van Harmelen, and Qing Hu. Constructing knowledge graphs of depression. In *International Conference on Health Information Science*, pages 149–161. Springer, 2017a.

Rong Xu, Li Li, and QuanQiu Wang. driskkb: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. *BMC bioinformatics*, 15(1):105, 2014b.

Yueyi I Liu, Paul H Wise, and Atul J Butte. The" etiome": identification and clustering of human disease etiological factors. In *BMC bioinformatics*, volume 10, page S14. BioMed Central, 2009.

Qing Zeng and James J Cimino. Automated knowledge extraction from the umls. In *Proceedings of the AMIA Symposium*, page 568. American Medical Informatics Association, 1998.

Elizabeth S Chen, George Hripcsak, Hua Xu, Marianthi Markatou, and Carol Friedman. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association*, 15(1):87–98, 2008.

Rong Xu and QuanQiu Wang. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC bioinformatics*, 14(1):181, 2013a.

Rong Xu and QuanQiu Wang. Toward creation of a cancer drug toxicity knowledge base: automatically extracting cancer drug—side effect relationships from the literature. *Journal of the American Medical Informatics Association*, 21(1):90–96, 2013b.

David A Hanauer, Mohammed Saeed, Kai Zheng, Qiaozhu Mei, Kerby Shedden, Alan R Aronson, and Naren Ramakrishnan. Applying metamap to medline for identifying novel associations in a large clinical dataset: a feasibility analysis. *Journal of the American Medical Informatics Association*, 21(5):925–937, 2014.

Ramakanth Kavuluru, Sifei Han, and Daniel Harris. Unsupervised extraction of diagnosis codes from emrs using knowledge-based and extractive

text summarization techniques. In *Canadian conference on artificial intelligence*, pages 77–88. Springer, 2013.

Tejaswi Rohit Anupindi and Padmini Srinivasan. Disease comorbidity linkages between medline and patient data. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 403–408. IEEE, 2017.

Cesar A Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas Alexander Christakis. A dynamic network approach for the study of human phenotypes. *Plos Computational Biology*, 5(4): e1000353, 2009.

Jean-Baptiste Escudié, Bastien Rance, Georgia Malamut, Sherine Khater, Anita Burgun, Christophe Cellier, and Anne-Sophie Jannot. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. *BMC medical informatics and decision making*, 17(1):140, 2017.

Di Zhao and Chunhua Weng. Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of biomedical informatics*, 44(5):859–868, 2011.

Min-Ling Zhang and Kun Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 999–1008. ACM, 2010.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333, 2011.

Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012.

Everton Alvares-Cherman, Jean Metz, and Maria Carolina Monard. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Systems with Applications*, 39(2):1647–1655, 2012.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089, 2010.

Yahong Zhang, Yujian Li, and Zhi Cai. Correlation-based pruning of dependent binary relevance models for multi-label classification. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 399–404. IEEE, 2015.

Elena Montañes, Robin Senge, Jose Barranquero, José Ramón Quevedo, Juan José del Coz, and Eyke Hüllermeier. Dependent binary relevance models for multi-label classification. *Pattern Recognition*, 47(3):1494–1508, 2014.

Abhishek Kumar, Shankar Vembu, Aditya Krishna Menon, and Charles Elkan. Beam search algorithms for multilabel learning. *Machine learning*, 92(1):65–89, 2013.

Abdulaziz Alali and Miroslav Kubat. Prudent: A pruned and confident stacking approach for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2480–2493, 2015.

Robin Senge, Juan José Del Coz, and Eyke Hüllermeier. On the problem of error propagation in classifier chains for multi-label classification. In *Data Analysis, Machine Learning and Knowledge Discovery*, pages 163–170. Springer, 2014.

Chunming Liu and Longbing Cao. A coupled k-nearest neighbor algorithm for multi-label classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 176–187. Springer, 2015.

Muhammad Atif Tahir, Josef Kittler, and Ahmed Bouridane. Multilabel classification using heterogeneous ensemble of multi-label classifiers. *Pattern Recognition Letters*, 33(5):513–523, 2012.

Ping Li, Hong Li, and Min Wu. Multi-label ensemble based on variable pairwise constraint projection. *Information Sciences*, 222:269–281, 2013.

Amirreza Mahdavi-Shahri, Mahboobeh Houshmand, Mahdi Yaghoobi, and Mehrdad Jalali. Applying an ensemble learning method for improving multi-label classification performance. In *2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–6. IEEE, 2016.

Runzhi Li, Wei Liu, Yusong Lin, Hongling Zhao, and Chaoyang Zhang. An ensemble multilabel classification for disease risk prediction. *Journal of healthcare engineering*, 2017(8051673):1–10, 2017.

Xiangnan Kong, Bokai Cao, and Philip S Yu. Multi-label classification by mining label and instance correlations from heterogeneous information networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM, 2013.

Yang Zhou and Ling Liu. Activity-edge centric multi-label classification for mining heterogeneous information networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1276–1285. ACM, 2014.

Lina F Soualmia, Saoussen Sakji, Catherine Letord, Laetitia Rollin, Philippe Massari, and Stéfan J Darmoni. Improving information retrieval with multiple health terminologies in a quality-controlled gateway. *Health information science and systems*, 1(1):8, 2013.

Yi-Ting Cheng, Yu-Feng Lin, Kuo-Hwa Chiang, and Vincent S Tseng. Mining sequential risk patterns from large-scale clinical databases for early assessment of chronic diseases: a case study on chronic obstructive pulmonary disease. *IEEE journal of biomedical and health informatics*, 21(2): 303–311, 2017.

Chu Yu Chin, Meng Yu Weng, Tzu Chieh Lin, Shyr Yuan Cheng, Yea Huei Kao Yang, and Vincent S Tseng. Mining disease risk patterns from nationwide clinical databases for the assessment of early rheumatoid arthritis risk. *PloS one*, 10(4):e0122508, 2015.

Mengting Wan, Yunbo Ouyang, Lance Kaplan, and Jiawei Han. Graph regularized meta-path based transductive regression in heterogeneous information network. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 918–926. SIAM, 2015.

Xiangnan Kong, Philip S Yu, Ying Ding, and David J Wild. Meta path-based collective classification in heterogeneous information networks. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1567–1571. ACM, 2012.

Jung-Woo Ha, Adrian Kim, Dongwon Kim, Jeonghee Kim, Jeong-Whun Kim, Jin Joo Park, and Borim Ryu. Predicting high-risk prognosis from diagnostic histories of adult disease patients via deep recurrent neural networks. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 394–399. IEEE, 2017.

XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature communications*, 5: 4212, 2014.

A Tsanas, MA Little, and PE McSharry. A methodology for the analysis of medical data. In *Handbook of Systems and Complexity in Health*, pages 113–125. Springer, 2013.

Silvia Oliveros Torres, Heather Eicher-Miller, Carol Boushey, David Ebert, and Ross Maciejewski. Applied visual analytics for exploring the national health and nutrition examination survey. In *2012 45th Hawaii International Conference on System Sciences*, pages 1855–1863. IEEE, 2012.

Leo Egghe and Loet Leydesdorff. The relation between pearson's correlation coefficient r and salton's cosine measure. *Journal of the American Society for information Science and Technology*, 60(5):1027–1036, 2009.

Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics*, 2(5):S4, 2011.

David Bowes, Tracy Hall, and David Gray. Comparing the performance of fault prediction models which report multiple performance measures: recomputing the confusion matrix. In *Proceedings of the 8th International Conference on Predictive Models in Software Engineering*, pages 109–118. ACM, 2012.

Karen Gardner, Beverly Sibthorpe, Mier Chan, Ginny Sargent, Michelle Dowden, and Daniel McAullay. Implementation of continuous quality improvement in aboriginal and torres strait islander primary health care in australia: a scoping systematic review. *BMC health services research*, 18 (1):541, 2018.

Wee Pheng Goh, Xiaohui Tao, Ji Zhang, and Jianming Yong. Decision support systems for adoption in dental clinics: a survey. *Knowledge-Based Systems*, 104:195–206, 2016.

Rong Xu and QuanQiu Wang. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC bioinformatics*, 14(1):181, 2013c.

Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: a structural analysis approach. *Acm Sigkdd Explorations Newsletter*, 14(2): 20–28, 2013.

Ming Ji, Jiawei Han, and Marina 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. ACM, 2011b.

Adler Perotte, Rajesh Ranganath, Jamie S Hirsch, David Blei, and Noémie Elhadad. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, 22(4):872–880, 2015.

Montserrat Cases, Laura I Furlong, Joan Albanell, Russ B Altman, Riccardo Bellazzi, Scott Boyer, Angela Brand, Anthony J Brookes, Søren Brunak, Timothy W Clark, et al. Improving data and knowledge management to better integrate health care and research. *Journal of internal medicine*, 274 (4):321–328, 2013.

Britta Böckmann and Katja Heiden. Extracting and transforming clinical guidelines into pathway models for different hospital information systems. *Health information science and systems*, 1(1):13, 2013.

Yuka Tateisi. Resources for assigning mesh ids to japanese medical terms. *Genomics & Informatics*, 17(2):e16, 2019.

João Pita Costa, Luka Stopar, Flavio Fuart, Marko Grobelnik, Raghu Santanam, Chenlu Sun, Paul Carlin, Michaela Black, and JG Wallace. Mining medline for the visualisation of a global perspective on biomedical knowledge. In *KDD 2018 (24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining)*, pages 1–2, 2018.

Thuan Pham, Xiaohui Tao, Ji Zhang, Jianming Yong, Wenping Zhang, and Yi Cai. Mining heterogeneous information graph for health status classification. In *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, pages 73–78. IEEE, 2018.

Supriya Supriya, Siuly Siuly, Hua Wang, Jinli Cao, and Yanchun Zhang. Weighted visibility graph with complex network features in the detection of epilepsy. *IEEE Access*, 4:6554–6566, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Guoqing Zheng and Jamie Callan. Learning to reweight terms with distributed representations. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 575–584. ACM, 2015a.

Guoqing Zheng and Jamie Callan. Learning to reweight terms with distributed representations. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 575–584. ACM, 2015b.

P Mohamed Shakeel, S Baskar, VR Sarma Dhulipala, and Mustafa Musa Jaber. Cloud based framework for diagnosis of diabetes mellitus using k-means clustering. *Health information science and systems*, 6(1):16, 2018.

Gang Luo. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Health information science and systems*, 4(1):2, 2016.

Svetla Boytcheva, Galia Angelova, Zhivko Angelov, and Dimitar Tcharaktchiev. Mining comorbidity patterns using retrospective analysis of big

collection of outpatient records. *Health information science and systems*, 5 (1):3, 2017.

Fuming Sun, Jinhui Tang, Haojie Li, Guo-Jun Qi, and Thomas S Huang. Multi-label image categorization with sparse factor representation. *IEEE Transactions on Image Processing*, 23(3):1028–1037, 2014.

Yong Luo, Dacheng Tao, Chang Xu, Dongchen Li, and Chao Xu. Vector-valued multi-view semi-supervsed learning for multi-label image classification. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, volume 2013, pages 647–653, 2013.

Anastasios Dimou, Grigorios Tsoumakas, Vasileios Mezaris, Ioannis Kompatsiaris, and Ioannis P Vlahavas. An empirical study of multi-label learning methods for video annotation. In *7th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 19–24, 2009.

Gulisong Nasierding, Yong Li, and Atul Sajjanhar. Robustness comparison of clustering—based vs. non-clustering multi-label classifications for image and video annotations. In *2015 8th International Congress on Image and Signal Processing (CISP)*, pages 691–696. IEEE, 2015.

Rui-Wei Zhao, Guo-Zheng Li, Jia-Ming Liu, and Xiao Wang. Clinical multi-label free text classification by exploiting disease label relation. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 311–315. IEEE, 2013.

Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. Large-scale multi-label text classification—revisiting

neural networks. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 437–452. Springer, 2014.

Yang Tao, Zhu Cui, and Zhu Wenjun. A multi-label text classification method based on labels vector fusion. In *2018 International Conference on Promising Electronic Technologies (ICPET)*, pages 80–85. IEEE, 2018.

Yuanjian Zhang, Duoqian Miao, Zhifei Zhang, Jianfeng Xu, and Sheng Luo. A three-way selective ensemble model for multi-label classification. *International Journal of Approximate Reasoning*, 103:394–413, 2018b.

Huawen Liu, Xindong Wu, and Shichao Zhang. Neighbor selection for multilabel classification. *Neurocomputing*, 182:187–196, 2016.

Zhiling Cai and William Zhu. Feature selection for multi-label classification using neighborhood preservation. *IEEE/CAA Journal of Automatica Sinica*, 5(1):320–330, 2017.

Wei-Jie Chen, Yuan-Hai Shao, Chun-Na Li, and Nai-Yang Deng. Mltsvm: a novel twin support vector machine to multi-label learning. *Pattern Recognition*, 52:61–74, 2016b.

Martin Ringsquandl, Steffen Lamparter, Ingo Thon, Raffaello Lepratti, and Peer Kröger. Knowledge graph constraints for multi-label graph classification. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 121–127. IEEE, 2016.

Jia Wu, Xingquan Zhu, Chengqi Zhang, and S Yu Philip. Bag constrained structure pattern mining for multi-graph classification. *IEEE transactions on knowledge and data engineering*, 26(10):2382–2396, 2014.

Xiangnan Kong and S Yu Philip. gmlc: a multi-label feature selection framework for graph classification. *Knowledge and Information Systems*, 31(2): 281–305, 2012.

Suping Xu, Xibei Yang, Hualong Yu, Dong-Jun Yu, Jingyu Yang, and Eric CC Tsang. Multi-label learning with label-specific feature reduction. *Knowledge-Based Systems*, 104:52–61, 2016.

Jun Huang, Guorong Li, Qingming Huang, and Xindong Wu. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE transactions on knowledge and data engineering*, 28(12):3309–3323, 2016.

Jianghong Ma and Tommy WS Chow. Label-specific feature selection and two-level label recovery for multi-label classification with missing labels. *Neural Networks*, 118:110–126, 2019.

Pengfei Zhu, Qian Xu, Qinghua Hu, Changqing Zhang, and Hong Zhao. Multi-label feature selection with missing labels. *Pattern Recognition*, 74: 488–502, 2018.

Jun Huang, Feng Qin, Xiao Zheng, Zekai Cheng, Zhixiang Yuan, Weigang Zhang, and Qingming Huang. Improving multi-label classification with missing labels by learning label-specific features. *Information Sciences*, 492: 124–146, 2019.

Runzhi Li, Hongling Zhao, Yusong Lin, Andrew Maxwell, and Chaoyang Zhang. Multi-label classification for intelligent health risk prediction. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 986–993. IEEE, 2016.

Xiaoqing Zhang, Hongling Zhao, Shuo Zhang, and Runzhi Li. A novel deep neural network model for multi-label chronic disease prediction. *Frontiers in genetics*, 10, 2019.

Felix G Rebitschek, Josef F Krems, and Georg Jahn. The diversity effect in diagnostic reasoning. *Memory & cognition*, 44(5):789–805, 2016.

Ludovic Dos Santos, Benjamin Piwowarski, and Patrick Gallinari. Multi-label classification on heterogeneous graphs with gaussian embeddings. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 606–622. Springer, 2016.

Jun Huang, Guorong Li, Shuhui Wang, Zhe Xue, and Qingming Huang. Multi-label classification by exploiting local positive and negative pair-wise label correlation. *Neurocomputing*, 257:164–174, 2017b.

Thuan Pham, Xiaohui Tao, Ji Zhang, Jianming Yong, Xujuan Zhou, and Raj Gururajan. Mekg: Building a medical knowledge graph by data mining from medline. In *International Conference on Brain Informatics*, pages 159–168. Springer, 2019.

Ziad Abdallah, Ali El-Zaart, and Mohamad Oueidat. An improvement of label powerset method based on priority label transformation. *International Journal of Applied Engineering Research*, 11(16):9079–9087, 2016.

Sanjay Yadav and Sanyam Shukla. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International conference on advanced computing (IACC)*, pages 78–83. IEEE, 2016.

Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3780–3788. JMLR. org, 2017.

A Santos, A Canuto, and Antonino Feitosa Neto. A comparative analysis of classification methods to multi-label tasks in different application domains. *Int. J. Comput. Inform. Syst. Indust. Manag. Appl*, 3:218–227, 2011.

Thuan Pham, Xiaohui Tao, Ji Zhang, and Jianming Yong. Constructing a knowledge-based heterogeneous information graph for medical health status classification. *Health Information Science and Systems*, 8(1):1–14, 2020.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2017.