The Islamic University–Gaza
Deanship of Research and Graduate Studies
Faculty of Information Technology
Master of Information Technology

الجامعـــــــة الإســـلاميـة – غـزة
عمادة البحث العلمي والدراسات العليا
كـليـــة تكنولوجيا المعلومات
ماجستير تكنولوجيا المعلومات

An Ontology Driven Data Mining Approach for CVD Diagnosing among Diabetic Patients in Gaza Strip

منهجية مستمدة من أنطولوجيا التنقيب عن البيانات لتشخيص أمراض القلب عند مرضى السكر في قطاع غزة

By:

Mohammed I. Qrenawi

Supervised by

Dr. Wael Al Sarraj

Assistant Professor of Computer Science

**A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Science in information Technology**

September 2018

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

An ontology driven data mining approach for CVD diagnosing among diabetic patients in Gaza Strip

تحديد منهجية مستمدة من أنطولوجيا التنقيب عن البيانات لتشخيص أمراض القلب عند مرضى السكر في قطاع غزة

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الاخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

I understand the nature of plagiarism, and I am aware of the University's policy on this.

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted by others elsewhere for any other degree or qualification.

Student's name:	محمد ابراهيم القريناوي	اسم الطالب:
Signature:		التوقيع:
Date:		التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية بغزة
The Islamic University of Gaza
عمادة البحث العلمي والدراسات العليا
هاتف داخلي: 1150

الرقم Ref ج.س.غ/35/.....
التاريخ Date 2018/10/14م

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة عمادة البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **محمد ابراهيم فرج القريناوي** لنيل درجة الماجستير في كلية **تكنولوجيا المعلومات**/

برنامج **تكنولوجيا المعلومات** وموضوعها:

منهجية مستمدة من انطولجيا التنقيب عن البيانات لتشخيص أمراض القلب عند مرضى السكر في قطاع غزة

An Ontology Driven Data Mining Approach for CVD Diagnosing Among Diabetic Patients in Gaza Strip

وبعد المناقشة التي تمت اليوم **الاحد 19 محرم 1440هـ** الموافق **2018/09/30م** الساعة **الثانية عشرة** مساءً، في قاعة مبنى **كلية** اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. وائل فكري السراج	**مشرفاً ورئيساً**	.....
د. توفيق سليمان برهوم	**مناقشاً داخلياً**	.....
د. محمد عوض عوض الله	**مناقشاً خارجياً**	.....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية تكنولوجيا المعلومات/برنامج تكنولوجيا المعلومات.**

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله تعالى ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق،،،

عميد البحث العلمي والدراسات العليا

أ. د. مازن إسماعيل هنية

P.O. Box 108, Rimal, Gaza, Palestine Tel: + 970 (8) 264 4400 هاتف: Fax: + 970 (8) 264 4800 فاكس: ص.ب. 108 الرمال، غزة، فلسطين
public@iugaza.edu.ps www.iugaza.edu.ps

الرقم العام للنسخة	3106853	اللغة	-

التاريخ: 20 / 10 / 20

الموضوع/ استلام النسخة الإلكترونية لرسالة علمية

قامت إدارة المكتبات بالجامعة الإسلامية باستلام **النسخة الإلكترونية** من رسالة

الطالب/ محمد ابراهيم خرج الله حمادي

رقم جامعي: 120130872 قسم: تكنولوجيا المعلومات كلية: تكنولوجيا المعلومات

وتم الاطلاع عليها، ومطابقتها بالنسخة الورقية للرسالة نفسها، ضمن المحددات المبينة أدناه:

- تم إجراء جميع التعديلات التي طلبتها لجنة المناقشة.
- تم توقيع المشرف/المشرفين على النسخة الورقية لاعتمادها كنسخة معدلة ونهائية.
- تم وضع ختم "عمادة الدراسات العليا" على النسخة الورقية لاعتماد توقيع المشرف/المشرفين.
- وجود جميع فصول الرسالة مجمَّعة في ملف (WORD) وآخر (PDF).
- وجود فهرس الرسالة، والملخصين باللغتين العربية والإنجليزية بملفات منفصلة (PDF +WORD)
- تطابق النص في كل صفحة ورقية مع النص في كل صفحة تقابلها في الصفحات الإلكترونية.
- تطابق التنسيق في جميع الصفحات (نوع وحجم الخط) بين النسخة الورقية والإلكترونية.

ملاحظة: ستقوم إدارة المكتبات بنشر هذه الرسالة كاملة بصيغة (PDF) على موقع المكتبة الإلكتروني.

والله ولي التوفيق،

311

توقيع الطالب

إدارة المكتبة المركزية

Abstract

Healthcare environment is rich of data, but still needs knowledge extraction that is necessarily important for saving people lives. Medical Knowledge discovery is a process of extracting knowledge patterns from biomedical data. It uses machine learning, statistical techniques and visualization to discover and present knowledge in a form that is easily comprehensible. Medical Knowledge discovery is useful and crucial for making effective decisions especially in developing strategies and policies of preventive medical treatments.

Data mining methods are the best-known way to recognize the hidden data standards. Ontology engineering used to improve knowledge domain representation. Ontology has been used to improve a specialized knowledge domains representation to get more efficiency in data mining results.

In this thesis, we apply ontology driven data mining techniques on a dataset of 300-diabetes patients form Gaza Strip in the year 2015-2016. These patients have cardiovascular disease. This process performed to identify the relationship between diabetes patients and their important laboratory tests specified by doctors. Doctors aim to investigate the probability of cardiovascular disease occurrence and stroke happening.

Rule induction, association rules and decision tree methods were used in support of ontology. In a late phase, we used frequent pattern discovery and rules induction method using ontological data mining algorithm (RMonto).

The findings of this study reveals that the use of ontologies minimizes the number of attributes in the pre-processing stage and helps in all data mining stages; in addition to its important role in ontological data mining method.

The results of data mining methods and ontological data mining shows that the impact of some laboratory tests like: LP(a),CRP,HDL,FBG,TG,LDH and Chol to predict CVD risk among T2DM patients with learning accuracy ratio about 91%.

Keywords: Knowledge discovery, cardiovascular, diseases, T2DM, Data mining, ontological data mining

الملخص

البيئة الصحية غنية بشكل كبير بالبيانات، التي بحاجة إلى استخراج المعرفة التي تعد ضرورية لإنقاذ حياة الناس. اكتشاف المعرفة الطبية هو عملية لاستخراج أنماط المعرفة Knowledge patterns من البيانات الطبية الحيوية. يستخدم التعلم الآلي والاحصاء والتصور لاكتشاف وتقديم المعرفة بشكل يسهل فهمه. اكتشاف المعرفة الطبية مفيد وحيوي لاتخاذ قرارات فعالة وخاصة في تطوير استراتيجيات وسياسات العلاجات الطبية الوقائية.

تعتبر أساليب التنقيب عن البيانات من أفضل الطرق المعروفة للتعرف على معايير البيانات المخفية. تستخدم هندسة علم الوجود Ontology Engineering في عملية تحسين تمثيل مجال المعرفة وفي تحسين تقنيات التنقيب على أساس الأنماط المكتشفة.

في هذه الرسالة، قام الباحث بتطبيق تقنيات قائمة على الأنطولوجيا Ontology على قاعدة بيانات تتعلق بمرضى السكري يعانون من الأمراض القلبية الوعائية. تتكون قاعدة البيانات من 300 مريض من قطاع غزة في العام 2015–2016، وذلك لتحديد العلاقة بين النوع الثاني من داء السكري والأمراض القلبية الوعائية من خلال التحاليل المختبرية الهامة التي يحددها الأطباء.

تم استخدام تقنيات تنقيب البيانات مثل قواعد الحث Induction، وقواعد الارتباط Association وطريقة شجرة القرار Decision Tree في هذا البحث بدعم من علم الوجود من أجل مقارنة النتائج مع نتائج اكتشاف نمط التكرار وطريقة الحث باستخدام قواعد استخراج البيانات الخوارزمية والأنطولوجية (RMonto).

تعكس نتائج هذه الدراسة أن استخدام الأنطولوجيات يقلل من عدد السمات في مرحلة ما قبل المعالجة ويساعد في جميع مراحل استخراج البيانات؛ بالإضافة إلى دورها الهام في تنقيب البيانات الأنطولوجية. تُظهر نتائج أساليب التنقيب عن البيانات وتنقيب البيانات في علم الوجود أهمية بعض الاختبارات المعملية مثل: LP (a) و CRP و HDL و FBG و TG و LDH و Chol للتنبؤ بمخاطر الأمراض القلبية الوعائية بين مرضى T2DM بنسبة تعلم عالية الدقة تقدر بـ91%

كلمات مفتاحية/ استكشاف المعرفة، التنقيب عن البيانات، علم الوجود، مرض السكري النوع الثاني، أمراض القلب والأوعية الدموية، تنقيب البيانات في علم الوجود

Dedication

I dedicate this work,

To my mother and father,

And to my wife and children.

Acknowledgment

I would like to express my sincere gratitude and thankfulness to my advisor Dr. Wael Al Sarraj, who provided me the opportunity to do such a research. I would also like to express my gratitude to Dr Tawfiq Barhom and Dr Mohammed Awadallah for accepting the invitation to discuss this work and enrich it with their valuable comments and observations. Moreover, my sincere thankfulness goes to my University professors, doctors and colleagues for their relentless support and assistance that they have provided me throughout my study year.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CVD	Cardiovascular Disease
CVDO	Cardiovascular Disease ontology
DL	Description Logic
DM	Data Mining
ES	Expert System
GT	Glossary of Terms
GUI	Graphical user Interface
KB	Knowledge Base
KDD	Knowledge Discovery Process
MOH	Ministry of Health
NDDG	National Diabetes Group
OWL	Online Writing Lab
RDF	Resource Description Framework
SPARQL	RDF query language
SVM	Support Vector Machine
SWRL	Semantic Web Rule Language
Type2DM	Diabetes mellitus
UNRWA	United Nations Relief and Works Agency
WWW	World Wide Web
XML	Extensible Mark-up Language
LPa	Lipoprotein little a
CRP	C-Reactive Protein
HDL	High-Density Lipoprotein
FBG	Fasting Blood Glucose
TG	Triglycerides
LDH	Lactate Dehydrogenase
Chol	Cholesterol
DE	Domain expert

Chapter 1

Introduction

Chapter 1
Introduction

1.1 Background and Context

Diabetes mellitus (normally known as diabetes) is an illness of the pancreas - an organ which is located in the human body behind the human stomach which produces the insulin hormone which helps the body digest food to generate energy. When we say that a person has diabetes, this means that the pancreas not able to produce enough insulin, or uses the insulin in a wrong way due to medical malfunction, or both. Insulin and glucose works together with bloodstream to help it enter the body's cells. The shortage of insulin lets the cells to starve and this causes glucose levels in the blood to rise up, creating a condition of high levels of sugar in the human body or diabetes (Badran & Laher, 2012). When glucose levels in the blood are out of control, it may increase the risk of heart diseases and stroke occurrences(Collaboration, 2010). People with type 2 diabetes may suffer of some dangerous conditions that contribute to increase the risk for developing cardiovascular disease like abnormal cholesterol and high triglycerides and Obesity (Collaboration, 2010). Cardiovascular diseases are a set of disorders of the heart, vessels and vascular diseases of the brain. Cardiovascular disease considered as the most dangerous disease that causes death and disability (Whelton et al., 2012). According to the statistics of Ministry of Health in Gaza strip (MOH, 2015). Figure 1.1 illustrates the prevalence of diabetes was 15.2 %; the prevalence of cardiovascular diseases was 27.4 % and prevalence of cardiovascular diseases and diabetic together 16.2%.

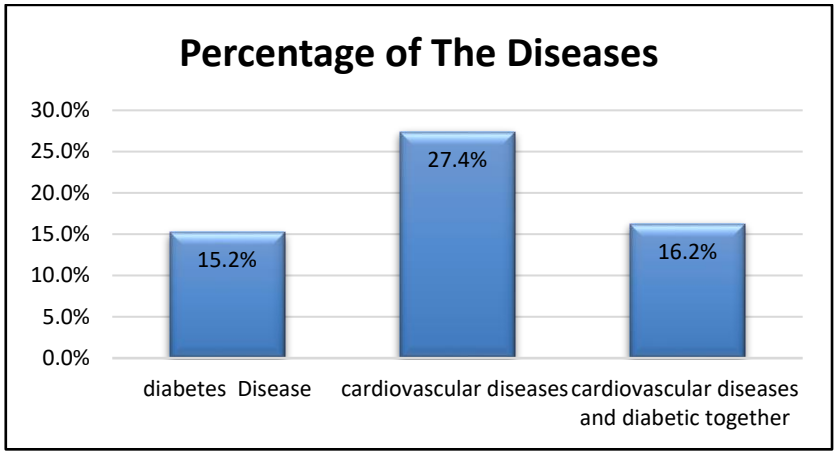


Figure 1.1 Prevalence of CVD and Diabetes in Gaza Strip in 2015(MOH, 2015)

Figure 1.2 shows the Sum of People with diabetes may reach 438 million in year 2030 when compared to 150 million diabetes patients in year 2000 (Ogurtsova et al., 2017). Diabetes decreases the normal life hope. The significant illness is linked with diabetes arises from microvascular complications, increased risk of microvascular complications (stroke, peripheral vascular disease, and ischemic heart disease), and endangers the quality of life (Packer, 2018).

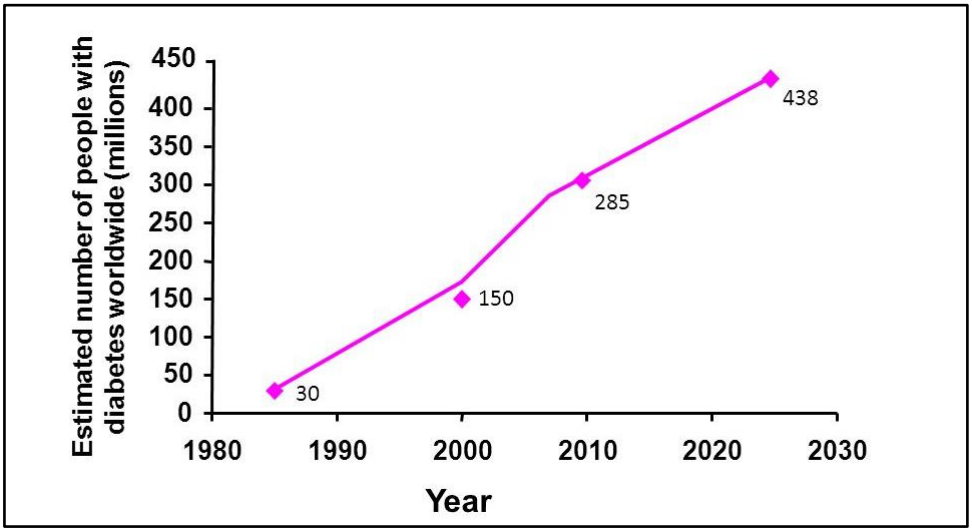


Figure 1.2 estimated Number of People with diabetes worldwide (Ogurtsova et al., 2017)

In the recent years, computer technologies are used widely for developing new strategies and techniques to diagnose diseases or improving health care, some of

these technologies are Knowledge Discovery/Data mining, Ontology, Expert Systems (ES) and other tools (Manogaran, Varatharajan, & Priyan, 2018).

Data mining is an analytical tool to analyse data and encapsulate it into useful information (Manogaran et al., 2018). It allows users to analyse data from many different perspectives and angels and summarizing it into useful information that can be used to increase revenue. Data mining is now widely used in many fields like image classification, Ranking and medical diagnosis (sick/healthy). In medical field, data mining algorithms used to mine the hidden knowledge in the dataset of the medical domain. The discovered knowledge and patterns may help in decision making and saving of lives (Kaushik et al., 2018).

Data mining can find out new healthcare knowledge for clinical and administrative decision-making as well as generate scientific hypotheses from experimental data, clinical databases, and/or biomedical literature (Ray, 2018). Data mining applications in healthcare areas such as the healthcare detection of fraud and abuse, appraisal of efficiency in treatment, hospital infection control, healthcare management, customer relationship management and identification of high-risk patients (Durairaj & Ranjani, 2013).

Ontology has been recognized as philosophical study of being in general(Gruber, 1993). Ontologies considered as acknowledge base or special kind of knowledge base. Which has its implicit conceptualization (Giaretta & Guarino, 1995). It offers answers to the customers without asking to help. In Artificial Intelligence community, ontology means a formal, explicit specification of a shared conceptualization where Conceptualization is an abstract model of some world phenomena. Ontology concepts and the relationship among those concepts should be explicitly defined. Further, ontology should be machine-readable and should capture consensual knowledge accepted by the community(Gruber, 1993).

Ontologies still work as an important tool for modelling the knowledge in different health care areas. This technology is useful because it allows implementing several kinds of interoperability and inference processes between different types of systems (Lavrač & Novak, 2013). In addition, it is a way of presenting the domain objects and the relation between these objects in a specific domain such as cardiovascular diseases. The basis of any decision support system is data repertories, where cleaned

data and integrated data for knowledge discovery stored through the data mining methods.

One need to select a data mining approach combined with ontology to enhance the overall data mining process since data mining process lack of semantic representation of data, lacking facilities in understanding users' mining intentions and generates unreasonable mining models. This approach benefits the mining process by providing intelligent assistance through the support of the ontologies(Wu, Lin, Jiang, & Wu, 2011).

Ontology with data mining algorithms bridges the semantic gap between the data, applications, data mining algorithms, and data mining results(Dou, Wang, & Liu, 2015). Many researches claims that there exists semantic shortage among data mining algorithms and the results of mining operation in all data mining stages. This shortage must be filled by human work, which clearly limits the efficiency of the overall process and the scope of data mining applicability(Dou et al., 2015).

Our research is concern with developing a Data Mining Based Ontology Approach and ontological data mining approach to diagnose and predict cardiovascular disease among diabetes patients. Such approaches will be helpful to utilize large volumes of available diabetes patient's related data, for extracting knowledge, such data like important laboratory tests, some of them is traditional for diabetes patients and conducted periodically and other new like High sensitive CRP, lipoprotein (a). In addition to this, symptoms diagnosis, treatments for the patients are collected and stored and some other needed information and expert's reviews.

1.2 Main Goal

The main goal of this research is developing an approach based on Ontological Data Mining Techniques to identify diabetes patients who are at risk to develop cardiovascular disease (CVD).

1.3 Specific objectives

- Build the domain knowledge for cardiovascular disease domain. Find solutions, and results

- Design and build ontology that describe and represent domain knowledge (Rules and Reasoning)
- Apply data mining based on ontology technique to discover the occurrence of cardiovascular diseases among diabetes patients.
- Apply ontological data mining algorithm to discover the occurrence of cardiovascular diseases among diabetes patients.
- Evaluate the research outcomes by asking experts and experiment some real cases in order to evaluate to correctness

1.4 Signification

In diabetes clinics of both UNRWA and Palestinian ministry of health in Gaza Strip, the practitioners use some traditional ways to diagnosis diabetic patients who are at risk to suffer from cardiovascular disease. Some important and effective tests not performed on the patients, they uses traditional laboratory tests and traditional ways, which may put the patient's life at risk.

Applying semantic web and data mining methods is a key approach to utilizing large volumes of available diabetes-related data for extracting knowledge. The severe of social impact of the specific disease requires submitting data mining and ontology in semantic web to be one of the main priorities in medical science research, which inevitably generates huge amounts of data and knowledge about the domain. Undoubtedly, therefore, ontology and data mining approaches are of great concern when it comes to diagnosis, management and other related clinical administration aspects. Hence, efforts were made to review the current literature on ontology and data mining approaches in diabetes and cardiovascular research

1.5 Scope and Limitations

1.5.1 Scope:

This research study will focus on diagnosis diabetic patients who are at risk to suffer from cardiovascular disease. The proposed methodology will let doctors, experts and others who concern to be familiar in cardiovascular disease domain, and easily diagnose patients who will suffer from cardiovascular disease before it happens and help others who are suffering from cardiovascular disease.

The proposed methodology will help doctors, patients, nursing to deal with diabetes patients in a correct way to avoid cardiovascular disease occurrence.

1.5.2 Limitations:

The proposed ontology for this research is a lightweight ontology. Lightweight ontology is an ontology with restricted expressiveness(Giunchiglia & Zaihrayeu, 2009). Lightweight ontology will be developed to suites the research objectives and based on proposed dataset

The dataset collected in Gaza Strip in year 2015-2016 and will not deal directly with X-Rays, CT and Echocardiogram; but the results of these images which written in the reports will be taken in consideration, so the accuracy of this data will depend on doctor's opinion.

The proposed methodology will mainly focus on patient's laboratory tests to diagnose the CVD.

This proposed approach will focus on diabetes patients only, since they are the most vulnerable to cardiovascular disease, according to statistics of ministry of health.

1.6 Methodology Overview

In order to improve the cardiovascular disease diagnosis, we need to build a model that includes ontology, data mining and ontological data mining using known algorithms and methods but in a new and novel way with different arrangement to accomplish the objectives of the research.

We intend to distribute the research work into five main steps as shown in Figure 1.3. The first step starts with CVD data collection. The second step concern with ontology building. The third step contains data mining algorithms application. In the fourth step, we applied ontological data mining algorithm. The five and last stage contains mapping the models form data mining an ontological data mining to get the final valid results.

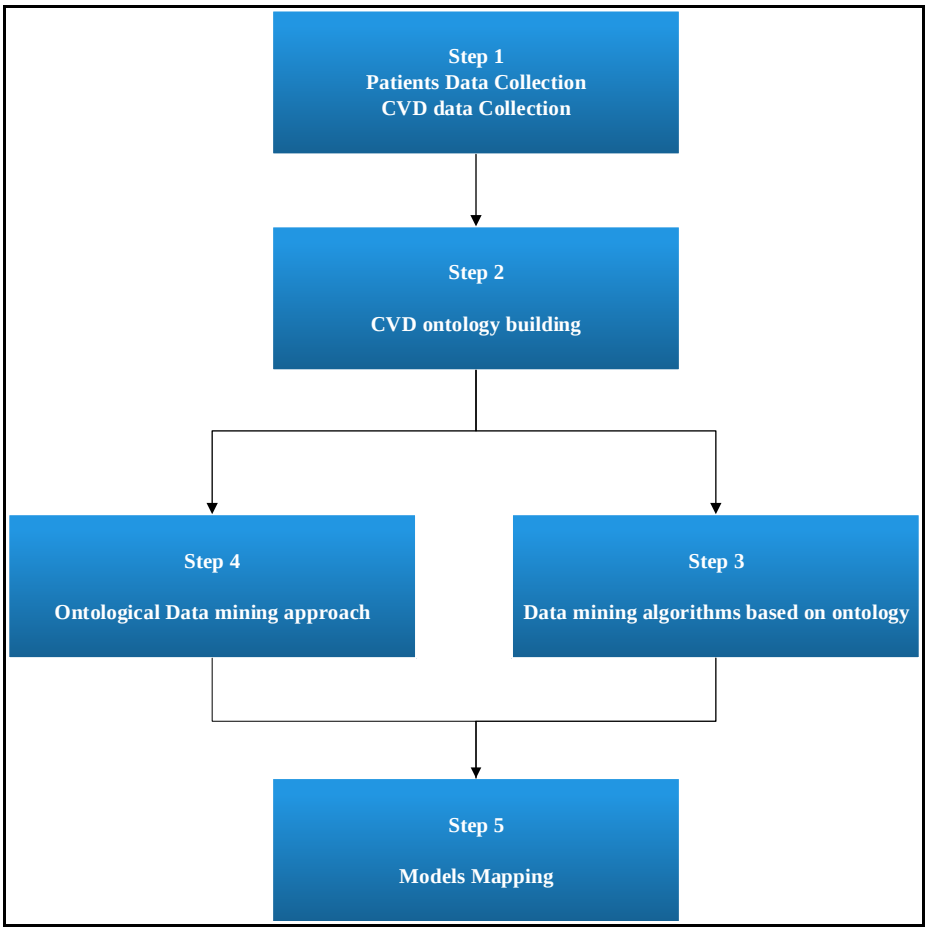


Figure 1.3 Proposed Methodology Steps

1.6.1 Step 1: Collecting patient information and laboratory tests

In this phase, Blood samples and important information will be collected. The collected data is a real world dataset. All tests are conducted in Gaza strip. It consists from 300 patient and other normal people. The data set contains about 40 data attributes about the patient. These attributes includes some personal information like name, age, smoking …etc. and other diagnosing information in addition to the most important part, which is laboratory tests results like FGB, LPa, CRP, Chol …etc. Collected data will be stored at relational database after preparing it according to ontology and remove all unaffected attributes to enhance the mining process and get results that are more accurate in the mining stage.

1.6.2 Step 2: Building ontology

METHONOLOGY methodology (Al-Hamadani & Alwan, 2015) will be used to develop an ontology because it has depicted design stages depth. Stages are interfere with each other and operates simultaneously.

The Ontology Component will be built using Protégé Framework to define and describe cardiovascular domain including classes, properties, attributes, and relations between domain entities.

Protégé is a flexible, user-friendly environment used to create and modify Ontology systems. It allows its users to add plug-ins, as they needed. It is supported with reasoner to check the consistency of the designed ontology and to infer new information base on the design of the existing system).Banaś, 2012 (

The main classes of the proposed ontology initially will be Patient, Diseases, Symptoms, Treatment, Patient-Info, Laboratory tests.

1.6.3 Step 3: Data mining based Ontology approach:

This approach is used to benefit from ontology inferences starting from using reasoner to refine and filters data which collected from hospitals. Ontology is the main role for data understanding and preprocessing stage. Ontologies can help a new miner to be familiar with most important concepts and relationships, it also allow to recognize not obvious assumptions. For example we can benefit from mapping elements of an ontology on elements of the data scheme and vice versa in the identification of missing attributes that should be added to the data set and also benefit in redundant attributes identification process.

1.6.4 Step 4: Ontological Data Mining approach:

Ontological data mining is a data mining approach where ontologies are used as background for domain knowledge(Ławrynowicz & Potoniec, 2014). This approach carry out the algorithm of pattern mining where pattern selection is routed throw a measure that takes the classification job in consideration. The proposed cardiovascular disease ontology which illustrated in Step 1 section 1.5.1 is considered as the knowledge base of this algorithm. This proposed ontology which

store all information about the diabetes patient who are suffering of CVDs is the data set to be mined. The set of features which are created (pattern set) is moved to the induction model algorithm to generate the final model (Ławrynowicz & Potoniec, 2014).

1.6.5 Step 5: Models mapping and Evaluation of proposed methods:

Different evaluation methods will be used to measure the value of the results. We will ask the expert about the generated rules and the new terms that inferred by the ontology reasoner and check the ontology consistency using pellet reasoner. We will compare the rules generated from data mining and ontological data mining to decide the best model for dealing with medical data and advice according to the high accuracy rules model.

Data mining results evaluation will be conducted by calculating the confidence, f-measure and accuracy ratio. In addition to that we will evaluate the model generated by ontological data mining using linear regression method.

Finally the results will be displayed to the cardiologist for evaluation and feedback.

1.7 Overview of the Thesis

This thesis organized as follow:

Chapter 1 Introduction: This chapter starts by presenting a general overview about the research, research objectives, and importance of the research, limitations, methodology which will be used in this research and the evaluation methods that will be conducted.

Chapter 2 State of the Art: This chapter presented the theoretical and technical foundations that are related to the research problem. We will survey the concepts of diabetes, cardiovascular disease, ontology, data mining and ontological data mining. Previous researches about data mining and ontology will be reviewed at its lowest abstraction level.

Chapter 3 Research proposal and Methodology: This chapter introduces the methodology of this research in details, tools and equipment used and the evaluation methods that will be conducted

Chapter 4 Cardiovascular Diseases Ontology Construction: This chapter introduces the empirical and theoretical steps to build and develop specialized cardiovascular disease ontology from scratch, using expert knowledge, reused ontologies related to the same domain and some medical documentations.

Chapter 5 Experimental Results: This chapter introduces the data mining experiments and ontological data mining experiments which are conducted on cardiovascular disease patient's data set. The mining algorithms used are association rules method, rule induction method and decision tree assisted by ontology design to get higher accuracy learning models. Frequent pattern discovery and rule induction methods are conducted to mine the knowledge base and get the affected factors on the patients.

Chapter 6 Recommendations and Evaluation: This chapter introduces the results of data mining methods and ontological data mining. Results are explained and evaluation methods are conducted.

Chapter 7 Conclusion and future work: this chapter will conclude the research and speaks about future works.

1.8 Summary

The preceding chapter have presented the best available knowledge about the thesis. It starts by presenting an introduction to the diabetes and CVD and terminologies related to it. Research problem and the importance of using ontology and data mining algorithm in diagnosing process are introduced. Research limitations and scope also are reviewed. The limitation is that the proposed ontology is lightweight not heavy weight ontology, due to the time needed to build such ontology and other listed limitations. The methodology of the research is presented and explained well

Chapter 2

Literature Review

Chapter 2

Literature Review

This chapter will survey the theoretical and technical foundations related to the research problem. Diabetes, cardiovascular disease, ontology, data mining and ontological data mining bases will be reviewed. The research related previous problem works will be discussed.

2.1 Background and literature review:

2.1.1 Diabetes Mellitus:

There are two fundamental types of Diabetes identified by the National Diabetes Data Group (NDDG), which are insulin-dependent (Type 1) and non-insulin-dependent (Type 2) (Association, 2017). Type 2 diabetes mellitus (T2DM) is a major cause of incapacity and loss of life global. Humans with T2DM have greater threat of cardiovascular sickness, stroke and early loss of life (Casey, 2018). The human is a sick of diabetes from type 2 when his body using insulin improperly. That is known as insulin resistance. At the beginning, your pancreas increases the production of insulin to compensate it but, over time it is not able to hold up and can't make sufficient insulin to maintain your blood glucose at normal stages. It observed that the type 2 diabetes mellitus disease is associated with cardiovascular Disease (CVD) (Larsson et al., 2018). The patient who suffering from diabetes is the most susceptible to be a CVD patient (Larsson et al., 2018). The diabetes cases risk of problem in some body parts some of these problems are(Agarwal, 2018):

- All types of heart disease
- Stroke occurrence
- Hypertension (High blood pressure)
- Eye damage and blindness
- Problems in hearing
- Kidney damage, Foot damage and even amputation

2.1.2 Cardiovascular disease:

CVD could be a terribly serious health problem that prevents the heart or blood vessels from working properly. once our heart and blood vessels are working in a good way or functioning normally, blood simply flows and is spread on the body organs freely. if our heart isn't pumping blood in a normal way, this means that there is a clog in the blood vessels, and the blood will not delivered to several vital elements of our body (Walters, 2008). Not having blood perpetually delivered throughout our body will cause serious health problem or perhaps death. Although some human are born with certain types of CVD, most of them develop CVD as a result of poor lifestyle habits, such as eating unhealthy foods, not obtaining enough exercise or victimization of smoking (Mozaffarian et al., 2016). A stroke happens once a vessel that carries oxygen to the brain gets blocked or bursts. When that happens, blood can’t deliver oxygen to part of the brain and that part of the brain starts to die. Without the correct medical treatments, a stroke will cause a brain damage. Those who have a stroke might have issues in speaking, seeing or moving commonly for a small time or for the remainder of their lives. The person will even die from a stroke. Older adults are more likely to have a stroke than younger people, but people can have a stroke at any age. alternative factors that increase risk of stroke embody high vital sign, fag smoking, diabetes, alternative kinds of upset, and high blood cholesterol. There are some of general symptoms for cardiovascular disease, these symptoms are:

- Sudden chest pain
- Effort chest pain
- Shortness of breath

2.1.2.1 Cardiovascular disease Diagnosing:

Doctors performs physical exams and reviewing the family history as a first stage in diagnosing process. Some blood tests will be conducted and other x-rays procedures in addition to ECG tests.

In general Cardiovascular Disease patient diagnosing methods listed as follow:

- Check heart rate
- Check blood pressure
- Physical exam
- Some laboratory tests is performed on a sample of blood like cholesterol and Triglycerides.

2.1.3 Semantic Web (SW):

Berners-Lee in 1966 suggested an idea to enrich the Web by machine understandable information which supports the user in his tasks, which called Semantic Web (SW) (Jain & Singh, 2013). The idea matured when he observed that a huge amount of data has a limited machine support and interpretable by humans only. The Semantic Web (SW) considered as an extension of the World Wide Web where documents are enriched by annotations in machine-processable language like OWL, extensible Markup Language (XML), Resource Description Framework (RDF), DAML (DARPA Agent Markup Language) (Jain & Singh, 2013). The semantic web aims to let the users find clear answers to their queried by maintaining the Semantic Web Inference Engine which responsible for results and queries between users and software agents. The results of semantic web which is tasks automation, led to more fast and accurate data processing and can drive to better utilization of data by sharing and reusing (Kolias, Stoitsis, Golemati, & Nikita, 2014). Semantic web simply adding logic to the World Wide Web, by expressing the meaning of data and the relations between the terms. There are many existing applications and researches in the medical field based on semantic web. To reach this goal the Semantic Web will be built up in different levels, the one we care about is ontologies.

2.1.4 Ontology:

Ontology is concept originates from old philosophy. From a long time, the philosophers studied the theory of objects and their ties(Welty, 2003). Ontological engineering is the activities that concern the process of ontology development, methodologies, techniques, languages etc. Ontologies became the more formalized

conceptual models used in artificial intelligence. According to Gruber, ontology is "the specification of conceptualizations, used to help programs and humans share knowledge."(Welty, 2003). And can be defined as "Formal, Explicit, and Specification of Shared Conceptualization"(Jain & Singh, 2013). From the above definition, ontology should be formal to be understandable by the machines to get better semantics of the World Wide Web and the knowledge included should be human readable for greater semantics. It should explicitly define constrains types. Ontology should be sharable and not restricted to anyone to provide a simplified and clear view of a specified domain. Conceptualization means points model of some phenomenon to characterize relevant concepts(Jain & Singh, 2013). Ontology increases the performance of Information retrieval. It deals with events occurrence, instances, and the relations between concepts in order to demonstrate background knowledge on Semantic level, which is set of semantic structures, includes concepts and relations instead of simple words used in dictionary(Jain & Singh, 2013). Formal knowledge representation languages have been developed for encoding the knowledge included in ontology, the most known one is description logic language. Description Logics (DLs) - knowledge representation formalisms that represent the knowledge of an application domain by defining the relevant concepts of the domain through the roles and relationships and then using these concepts to specify properties of individuals occurring in the domain.

2.1.4.1 Ontology Components:

- Classes: concepts, abstract or specific. The main component of ontologies. Classes in an ontology should be organized in taxonomies
- Relationships: association between domain concepts. Describes the relation among individuals and how they are related to each other's. It can be expressed between individuals (A CVD has symptoms) or expressed between concepts (this Disease has symptoms a symptom).
- Functions: is a special type of relationship in which one of the relationship's elements is the result of a formula.

- Axioms: used for modeling sentences that are true. They represent knowledge that cannot be formally defined with the rest of the terms. Should be used to preserve consistency(Gómez-Pérez, Fernández-López, & Corcho, 2006).
- Instances: particulars which ontology describes or possibly could describe. Used for representing elements or individuals of an ontology. Instances are the base unit of an ontology that characterizing the entity of interest
- Properties and their values of the above components

2.1.4.2 Ontology Design principles:

- Clarity: communicating the importance of terms clearly. These terms should be language independent.
- Extensibility: Expect the use of shared vocabularies.
- Coherence: The achieved inferences should be Compatible with ontology definitions.
- Minimal ontological compromise: Compromises should be kept to a minimum, but guaranteeing the essentials(Gómez-Pérez et al., 2006).

Other principles:

- Classes and their subclasses should be well defined with
- Disjunctive and exhaustive knowledge.
- Name standardization

2.1.4.3 Ontology Building Methodologies

In general, methodologies provide you with a group of guidelines of performing the activities known within the ontology development method(Fernández-López, Gómez-Pérez, & Juristo, 1997). Ontology building will be difficult when the expert or the oncologist has no background knowledge on ontology engineering techniques. According to that the researchers proposed several mature ways to develop and build ontologies. A lot of successful applications in many fields explained the success of these methodologies(Brusa, Caliusco, & Chiotti, 2006). METHONTOLOGY methodology is considered one of the best and comprehensive ways to develop an

ontology in ontology engineering. This methodology depicted the design stages depth, which are interfere with each other and operates simultaneously (Al-Hamadani & Alwan, 2015; Fernández-López et al., 1997).

At this section we will explain the process and stages of METHONTOLOGY to develop and built ontology.
The process of this structured method for developing ontology as follow (Fernández-López et al., 1997):

2.1.4.4 Specification stage

The objective of this stage is to produce specification document which is either formal, informal or semi-formal using representations or competency questions. This stage will define the purpose and the scope of the proposed ontology. Before this stage end we must have answered the following questions: what domain will we cover? What reason ontology will be used? What type answers should be provided؟ Who will use and preserve the ontology?

2.1.4.5 Knowledge acquisition stage

The objective of this stage is to construct a glossary of terms. Knowledge Capturing is an independent part in the process of ontology development, but it is synchronized with other processes. Generality of knowledge acquisition done together with the requirements specification phase and become smaller through the progress of ontology development process. There are several sources of Knowledge at this stage: interviews, texts analysis which is formal and informal to determine the main structure of the ontology, and brainstorming which named conjunction techniques.

2.1.4.6 Conceptualization stage

This stage depends on the previous stage output. The first task starts with constructing a glossary of terms (GT). Terms include concepts, instances, verbs and properties used in designing the ontology (Gómez-Pérez et al., 2006). Every term that will be part of the ontology should include a brief definition such as synonyms

and acronyms، their type etc. which is useful in class hierarchy building and generating conceptualization

Table, in addition to that it is useful to determine the relations between classes. The second task is construct taxonomies in order to classify a set concepts and verbs. Then concepts classification tree and a verbs diagram is built for each set of related concepts and related verbs.

2.1.4.7 Integration stage

Ontology reusability stage to speed up ontology construction. This stage take into account the utilization of definitions of previously constructed ontologies instead of starting from the beginning. It's advisable to select ontologies that suitable to your conceptualization to be sure that the new and reused sets of taxonomies are based on the same group.

2.1.4.8 Implementation stage

At this stage, computer software's and tools are used to implements ontology. These tools is applicable in all design stages. Protégé is the most famous and reliable ontology editor with an intuitive interface , which allows the users to build their domain ontology easily (Youn & McLeod, 2006). Protégé software has translators and analyzers to check the errors in syntax and lexical.

2.1.4.9 Evaluation stage

This stage interferes with all other processing stages. The purpose of evaluation stage is to guarantee ontology reliability and consistency and its power in retrieving queries. Once this stage done, correct and valid knowledge base will provided to the user. Different factors are considered in ontology evaluation stage: firstly we have to compare ontology with others and choose the easier and suitable one, then check the size of the proposed ontology, the content, and simplicity (Lovrencic & Cubrilo, 2008). The domain expert has an important role in evaluation stage. Interviews with those experts will be conducted to check the validity of knowledge captured and tables generated and make sure the proposed ontology meets its purpose.

2.1.4.10 Documentation stage

Recommends to starts write guidelines in the method of building the ontology, to be used in publishing. This stage is conducted parallel with all design stages to write the need information about each stage.

2.1.5 Data Mining:

The goal of data mining process is discovering unknown implicit information from data sets and detects patterns in these set of data automatically. The useful generated pattern, if discovered will be generalized for predictions issues on future data. Different data mining algorithms used in data mining techniques and methods. Classification, prediction, association and clustering are the most known techniques which provides an explicit structural description (Gorskis & Chizhov, 2012).

Knowledge discovery process in databases are iterative methods, which found to help in discovering the hidden and meaningful patterns in data(Ilayaraja & Meyyappan, 2015). As shown and explained in Table 2.1 the process started by determination of goals and collecting data to evaluation and knowledge representation.

Table 2-1 Knowledge Discovery Process in Databases

notes	Phase	Description
May take 60%-80% of effort!	Determine Goals	Define the goal and the objectives according to the collected data
	Data understanding	Take a closer look at the data available for mining
	Data pre-processing	Data cleaning, data integration, data reduction, data transformation
Discovering patterns	Data mining	Choosing appropriate data mining function. Decision tree, rule induction, association rules
Is it Novel?	Evaluation	Dose the model achieve the goals, using evaluation methods.
Putting the results into use	Knowledge representation	visualization, transformation, removing redundant patterns, etc.

2.1.5.1 Association Rules:

Probabilistic in nature. Concern with finding the associations relationships in data sets. It shows the conditions which repeatedly occur in the given data. The provided rules takes the shape (if-then). The rules has value which express uncertainty degree

named support which expressed as a percentage. The support percentage include the count of transactions of all antecedent and consequent parts of the rule. Confidence value associated with the rule is the ratio of all items that included in transaction(Rajak & Gupta, 2008).

Support and confidence is an important measures. The rule with low support means that it may happen by chance. The confidence measures the inferences made by this rule. The formal definitions of these measurements are (id Items):

Association Rule (AR): implication $X \Rightarrow Y$ where $X,Y \subseteq I$ and $X \cap Y = \phi$;

Support of AR (s) $X \Rightarrow Y$: Percentage of transactions that contain $X \cup Y$

Confidence of AR (a) $X \Rightarrow Y$: Ratio of number of transactions that contain $X \cup Y$ to the number that contain X

2.1.5.2 Rule Induction:

Rule induction method which is considered one of the most significant method in machine learning algorithms. The hidden regularities in data are frequently used explicated in terms of rules(Rajak & Gupta, 2008).

Rule induction is considered one of the main tools for data mining all the time. Usually rules are expressions of the form If (attribute − 1, value − 1) and (attribute − 2, value − 2) and (attribute − n, value − n) then (decision, value)

2.1.5.3 Decision Tree:

Classifier used in machine learning and decision supports systems. It has the ability to produce a set of simple decisions by breaking down complex decision using divide-conquer technique, thus makes these solutions understandable and easier. As shown at Figure 2.1 the decision tree consists of finite number of nodes started by root - the most effectible one - node and some other nodes called leaves. Each internal node divide the instance space to two or more spaces according to a particular function (Rokach & Maimon, 2008; Seema, 2012). Decision tree construction involves the three phases: the construction Phase based on the entire

data set. Pruning phase which improves the performance by removing unnecessary branches and nodes. Processing phase which improves the understandability.

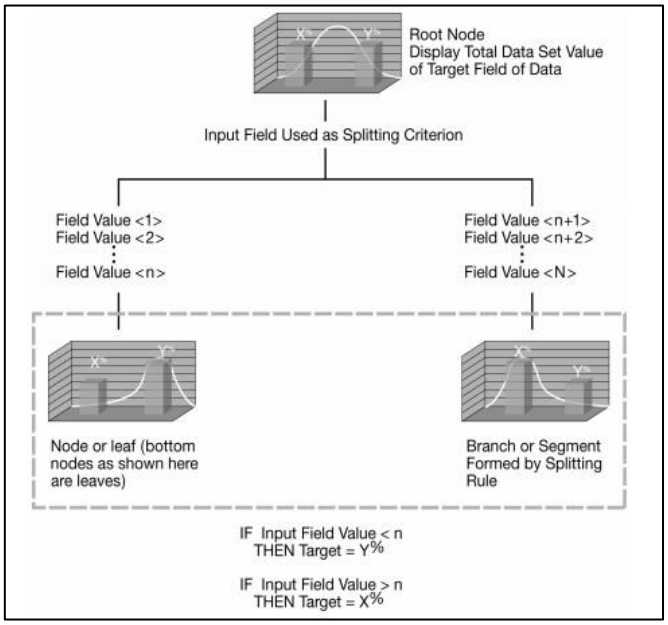


Figure 2.1 illustration of the Decision Tree (Seema, 2012)

As shown in Figure 2.2, Decision tree algorithms are composes the decision automatically from data set. The purpose of these algorithms to find optimal tree with the minimum number of nodes and leaves and construct the decision.

Decision tree offers may benefits in data mining. It's self-explanatory, handle different types of data, can process data set with some missing values, and its highly predictive model with high performance(Seema, 2012).

```
Algorithm:
Input: T// Decision Tree
       D// Input Database
Output: M// Model Prediction
DT Proc algorithm:
// simplest algorithm to illustrate prediction technique using DT.
For each t ∈ D do
n = root node of T;
While n not leaf node do
Obtain answer to question on n applied to t;
Indentify arc from t, which contains correct answer;
n = node at end of this arc;
Make prediction for t based on label of n;
```

Figure 2.2 Decision Tree Algorithm(Seema, 2012)

2.1.5.4 Performance Evaluation:

Confusion matrix:

A confusion matrix is the main source of calculating the accuracy of classification. It shows how many instances assigned to the target class (Dangare & Apte, 2012). Figure 2.3 shows the parameters of confusion matrix.

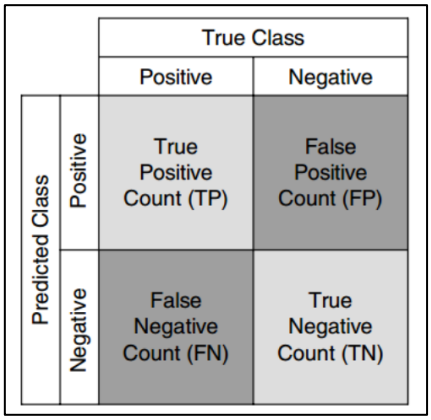


Figure 2.3 Confusion matrix

TP: records identified as true and they actually true.

FN: records identified as false and they actually true.

FP: records identified as true and they actually false.

TN: records identified as false and they actually false.

From the confusion matrix we can calculate (Dangare & Apte, 2012):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Eq. 2.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Eq. 2.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{Eq. 2.3}$$

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad \text{Eq. 2.4}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \quad \text{Eq. 2.5}$$

And finally the F-measure which is the weighted average between precision and recall

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad \text{Eq. 2.6}$$

2.1.6 Ontological Data Mining:

The concept ontological data-mining lead us to the third generation data mining approach, this approach uses extracted information from data in terms of ontology instead of empirical data. Ontological data mining approach depends on analysis of ontology for specific domain and analysis the information in data which can be explained in ontology terms. This approach has the ability to explain data mining results in ontology terms(Kreinovich, 2017).

2.1.6.1 Role of Ontological Data Mining:

Normally data preprocessing needs several understanding of the domain especially when you mine clinical datasets(Kuo, Lonie, Sonenberg, & Paizis, 2007). The miner who has no previous knowledge about the medical domain or any other domain, will

not be able to know which variable should be used when there is duplication in data or know the low important attribute to ignore it. In addition to that understanding of the data mining results always depends on domain understanding and domain knowledge to select features properly (Kuo et al., 2007). Ontology provide alternate knowledge source to captures concepts and relations which are collected from places and warehouses like dictionaries, internet, and domain experts. Ontology with data mining algorithms bridges the semantic gap between the data, applications, data mining algorithms, and data mining results(Dou et al., 2015). Many researches claims that there exists semantic shortage among data mining algorithms and the results of all mining operation in all data mining stages. This shortage must be filled by human work which clearly limits the efficiency of the overall process and limits the scope of data mining applicability. For example data mining can reveal that Triglycerides (TG) laboratory test is often accompanied by Cholesterol lab test but it cannot hypothesize that the patient is probably has cardiovascular disease. In addition to that data mining designed to be used on data from various domains, however it lacks the ability to determine semantics across different domains. Semantic web ontology is useful in specifying the semantics of any domain, it reduces the semantics gaps by annotating the data with rich semantics. Annotation assigns basic element of information links to formal semantic descriptions(Domingos, 2007; Dou et al., 2015). Ontology introduces prior knowledge about the domain which is useful in assisting data mining process. Prior knowledge guide the data mining operation from starts to end and effects all data mining stages. The united of ontology and graph representation as priori knowledge representing the distances between terms and concepts(Liu, Dou, Jin, Lependu, & Shah, 2013).

2.1.6.2 Fr-ONT-Qu algorithm

Based on pattern mining. Named Fr-ONT-Qu to maintain some continuity with the Fr-ONT algorithm. Fr-ONT-Qu designed to mine patterns represented as SPARQL queries over RDFS (Ławrynowicz & Potoniec, 2014). Fr-ONT-Qu discovers frequent/good patterns in knowledge base of any domain in the form of SPARQL patterns. The search space of the algorithm is restricted only to the parts specified on

the ontology classes. There are further limitation done, by setting the maximal complexity of the patterns(Ławrynowicz & Potoniec, 2014). Selecting the best pattern can be done in both supervised and unsupervised learning manners.

2.2 Related work:

In this section we will represent previous studies which are related to our research. The related work and studies are distributed on three main categories ontology, data mining and ontological data mining approaches.

2.2.1 Disease ontology Construction:

Gedzelman, Simonet and others (Gedzelman, Simonet, Bernhard, Diallo, & Palmer, 2005), introduces Ontology of cardiovascular diseases for concept-based information retrieval. The cardiovascular diseases ontology generated for NOESIS project which is an integrated project for European framework program in order to help in purposes of indexing and users support to assists in information retrieving through direct and indirect ways. The authors used an automated way from MeSH which refers to Medical Subject Headings and other existing resources. They used MOB API which is a GUI tool for the automation of extracting information. The extracted ontology contains 690 OWL classes, 2070 English expressions and 966 French expressions. Different models can interact with the constructed ontology. His ontology used in Noesis project for knowledge management and proposed a valid rules, and the texts used are automatically indexed by indexer platform. The knowledge generated from the ontology are shared using an annotation tool.

Al Hamadani and Alwan (Al-Hamadani & Alwan, 2015), introduces an ontology-based expert system for general practitioners to diagnose cardiovascular diseases. The designed ontology concern with diagnosing any kind of coronary artery diseases, which is type of cardiovascular disease. METHONOLOGY methodology used to construct the ontology of the disease, which is the most appropriate way in building ontology according to the author's notes. The purpose of constructing ontology to be

the knowledge base of the proposed system which is the core of any expert system. Knowledge base is the base of the proposed model which includes rule base and fact base which extracted from user interface module. SWRL rules are used to infer the appropriate medicine and the needed operation to the patient. The inferences depends on the knowledge base module rules and facts to reason the required decision. The results of the proposed expert system was 100% correct and precision was 0.872, while the recall was 0.831.

El-Sappagh and Ali (El-Sappagh & Ali, 2016) introduces DDO which refers to a diabetes mellitus diagnosis ontology. Based on standard ontologies the authors constructs a diabetes ontology for decision supports issues, and for enhancing knowledge sharing, and reusability and consistency. Existed ontology related to diabetes and it's all related to it are taken in consideration. The approach used in the construction of DOD was top-down and bottom-up to catalog relationships and attributes related. The methodology for constructing DOD consists of sequential 5 phases which are: Terms Determination, Ontologies reuse, Merging with OGMS[1], ontology coverage check, and ontology implementation. The proposed ontology containes more than 6400 concepts, about 48 properties, 13,551 annotations, and over 27,000 axioms were included in DDO. This ontology is implemented in the OWL 2 using Protégé tool. Comprehensive ontology which contains all diabetes-related involvement. It can serve as a base for diabetes knowledge and generates automatic reasoning.

Barton wt. al (Barton, Rosier, Burgun, & Ethier, 2014) introduces the cardiovascular disease ontology CVDO in OWL ontology language format to fulfilling the principles of OBO[2] foundry. The CVDO uses automatic reasoning the elaborate classification of the disease and derives inferences related to the disease classifications. CVDO reused high numbers of classes from OBO ontologies foundry and Makes important ontological unique characteristics between entities which are not well-identified in common medical language, in order to produce and elaborate classifications for the disease by automatic reasoning, and to derive interesting

[1] Refers to standard medical terminologies or ontologies
[2] Open Biological and Biomedical Ontology

inferences concerning disease classifications. CVDO has about 193 defined classes and is consistent with pellet reasoner.

2.2.2 Data Mining for Diseases diagnosing:

Parthiban and Srivatsa (Parthiban & Srivatsa, 2012), introduces machine learning methods in diagnosing heart disease for diabetic patients. The method concern with constructing a machine-learning methods model for prediction of heart disease occurrence among diabetic patients using Naïve bayes and Support Vector Machines methods depending on some diabetic diagnosis attributes. Dataset collected from health care institution of diabetic patients which have a database of patients with heart disease symptoms stored. The dataset has some personal information like Sex, age, weight and history and other medical related to heart disease diagnosing like blood pressure, last 4 months sugar level and total cholesterol level. KDD process conducted, and attribute normalization to convert attributes to numeric which is required in SVM. Naïve bayes model was able to classify 74% of the input instances correctly. Support Vector Machines method SVM's have proven to be a classification technique with excellent predictive performance, with accuracy about 94%.

Ilayaraja and Meyyappan (Ilayaraja & Meyyappan, 2015), introduces an efficient data mining method to predict the risk of heart diseases through frequent item sets. The authors proposed an efficient way that predict the dangerous level of the patients who are having heart disease through heart disease patient data sets. The proposed method is applicable to any different dataset in medicine. The heart disease patient dataset contains 1000 patient with 19 symptoms of the disease like chest paint, shortness of breath …etc. They proposed new mining algorithm that neglects the generation of unwanted items, like rows with zero value and the item not satisfy the support number. This feature will save the execution time. The author compared the new method with old known methods like association rules and others the results was encouraging and the new method better than the existing one. The proposed method identified the risk level of patients efficiently, and this helps in saving life and diagnosing accurately.

Masethe wt. al. (Masethe & Masethe, 2014) introduces Prediction of heart disease using classification algorithms. The authors used various data mining algorithms، such as decision tree, J48, Naïve Bayes, REPTREE, CART, and Bayes Net to predict the heart attacks and to compare the best method of prediction. 11 attributes was taken in consideration in prediction process from a dataset which was collected from doctors clinics in South Africa. The target dataset attributes contains medical information like laboratory tests and symptoms and other personal information. The model was able to predict about 89 cases with positive risk factor to be attacked by heart disease. They found that there is no difference in the prediction when they used different classification algorithms. The changes was to help the physicians to predict risky cases in the practice and advise accordingly. The result shows prediction accuracy of 99%. The author advice about the usefulness of data mining algorithms in medical diagnosing.

2.2.3 Ontology and data mining:

Kazi and Kurian (Atiya Kazi, 2014) introduces an ontology based approach to data mining. The paper concentrates on mechanism of data mining and ontology cooperation during knowledge discovery process. It also discusses the problem of assessing certain application criteria using ontology, by deciding which ontologies are suitable for the current domain. Two phases formulating the methodology of this work. The knowledge discovery phase including its normal process and ontology building phase as an output of knowledge discovery process (data mining). Data mining phase starts with preprocessing process, data mapper and data mining algorithms. The proposed model parsing the results of mining process and constructs an ontology in OWL and XML. The goal of constructing ontology from mining process is to gain useful knowledge fully of semantics which is the power of ontology. Then use expert systems fabrications in knowledge management applications.

Branquinho wt. al.(Lucélia P. Branquinho, 2015), introduces Ontologies in support of data mining in a diagnostic medicine company based on association rules.

Diagnosing viral hepatitis ontology was constructed based on dataset of laboratory tests related to the disease. The laboratory tests added to allow generalization of mined attributes to suitable for any laboratory in data mining. Data set collected from diagnostic medicine company's database for mining process, contains 439 various laboratory tests. Ontology reasoner used to enhance the generalization operation (refinement and filtering) of the collected laboratory tests. Jena tool used for ontology construction and rules inference. Using ontology inferences construction throw the pre-mining phase reduces the number of attributes needed to be mined by Apriori algorithm and reduces the number of rules generated by the algorithm. The finding of this study reflects the importance of extracted patterns by identifying similarity between terms extremely depends on ontology development.

Gottgtroy wt. al. (Gottgtroy, Kasabov, & MacDonell, 2004), introduces an ontology driven approach for knowledge discovery in biomedicine. They analyzes the important rule of ontology combined with knowledge discovery process in knowledge extraction from Biomedicine filed. The authors found new methodology for discovering semantics to improve data mining process using ontology. The study explained how to build multidimensional ontology, which capable of sharing knowledge form different research communities. Hybrid approach of data mining and ontology defines a frame work of knowledge discovery driven by ontology in order to improve knowledge discovery in data sets. The frame work methodology contains five steps which are Ontology preparation, population, Instance selection, mining, and Ontology refining. The frame work determined the necessary steps that guides the selection of suitable techniques to improve ontology engineering for knowledge discovery.

Gorskis and Chizhov (Gorskis & Chizhov, 2012) introduces ontology building using data mining techniques. The authors explains the significance of elaborating the main concepts of data mining and the ontology to comprehend the connection between data mining and ontology structure. Classification and clustering techniques are used for ontology building. Data mining techniques used during ontology building ends with the classes of ontology as an instances. Ontology relations and structure which

are complicated, lets us to realize the information needed by expert system. The authors found out how data mining techniques could help in ontology building.

2.2.4 Related Work Conclusion:

The previous studies that related to our representation are explained in this chapter. We explained the benefit from these studies and its differences with our research methodology as follow:

1. **Ontology construction studies:**

   In section 2.2.1, we have reviewed the methodologies that used to build ontologies which related to the medical field and the reasons for building such ontologies. Ontologies are developed for information retrieval issues, building an expert system and for knowledge discovery and sharing issues. We chooses the best methodology of building ontologies according to the results of pervious works and authors recommendations(Al-Hamadani & Alwan, 2015) in addition to our research needs. Ontologies which are related to our domain are reused and helped in exploring terms and taxonomies. These ontologies and extracted terms are used to develop our CVD ontology from scratch.

   In our work we will design a general cardiovascular ontology not specific type of the disease. Our ontology is lightweight, describes the disease generally(Giunchiglia & Zaihrayeu, 2009). The proposed ontology will concern on adding all laboratory test that related to the cardiovascular disease domain according to the expert recommendations. Other rules will be added to the ontology after finishing the mining process. These rules will be resulted from data mining process to complete the terms defining the cardiovascular disease. Our ontology mainly will be used for helping data mining process and ontological data mining in addition to its rule in sharing and reusing

2. **Data mining studies:**

In section 2.2.2, we have reviewed the data mining methods which used in disease diagnosing and prediction. Studding these works helps in discovering the best way in diagnosing and prediction. Knowledge discovery process are explained and reused with some enhancement in our work. In addition to that we studied different data mining algorithms used to diagnose disease and benefit from it. Detecting out layer values and null in addition to conducting different evolution methods to evaluate the models. We used some of these methods in our work and could justify the best method.

In our work we will use ontology driven the knowledge discovery process. Ontology will be the main rule in understanding the domain and in preprocessing stage. We can benefit from the ontology in identifying the missing attributes and vice versa. Ontology can add value to domain understanding by identifying the requirements of CVD diagnosing and the laboratory tests needed. As will as it has a rule in explanation of results to the miner who are not a cardiologist. Our data set is real data and contains some new tests not conducted in Gaza Strip like lipoprotein (a) LPa and high sensitive CRP, to find the effects of those tests result on CVD occurrence.

3. **Ontology and Data mining studies:**

In section 2.2.3, we have reviewed the usage of ontologies and data mining together. This section concentrates on the cooperation of ontology and data mining in knowledge discovery and the benefit from this combination in producing fully of semantics knowledge. Some of the researches explains the rules of data mining in ontology building. We benefit from these studies to know the rule of ontology in data mining and how they can enhance the knowledge discovery process.

In Our work we used ontology driven data mining and ontological data mining methods in medical field. Our methodology used ontology in data mining process to enhance the mining results, since it helps in data understanding stage, data preparation stage and results explanations and understanding. In addition to use a pattern mining algorithm, to mine the

ontology itself which contains the data set. The used algorithm is assisted by ontology reasoner to enhance the learning process.

Our work uses different and novel methodology to enhance the results of the knowledge discovery process and to enhance the resulted knowledge base.

2.3Summary:

The preceding chapter have presented the theoretical background and the previous works which related to our work. This chapter is divided to two sections. The first section illustrated an overview about diabetes and cardiovascular disease and diagnosing issues. It also provide theoretical information about semantic web and ontology building process in addition to ontological data mining. In the second section, we studied different related works that use ontology, data mining and ontological data mining. These works concentrate on disease diagnosis using ontology and data mining and using ontology with data mining. We have reviewed the works that related to our representation and discussed the solutions and show how our solution represents the best methodology

In the next chapter we will explain the methodology of our research and the data set used and the used algorithms, tools and evaluation methods.

Chapter 3

Research proposal and Methodology

Chapter 3
Research proposal and Methodology

This chapter will present and explain the proposed methodology for identification of cardiovascular disease among diabetes patients.

3.1 Methodology steps:

As shown in Figure 3.1, methodology starts by collecting all available information about the cardiovascular disease from doctors and other resources, in addition to collect data set of patients in parallel, in order to starts ontology construction process and prepare the data set for mining process. The later step is ontology building process, then data mining process and ontological data mining to generates the models. Evaluation process will be conducted for ontology and generated models.

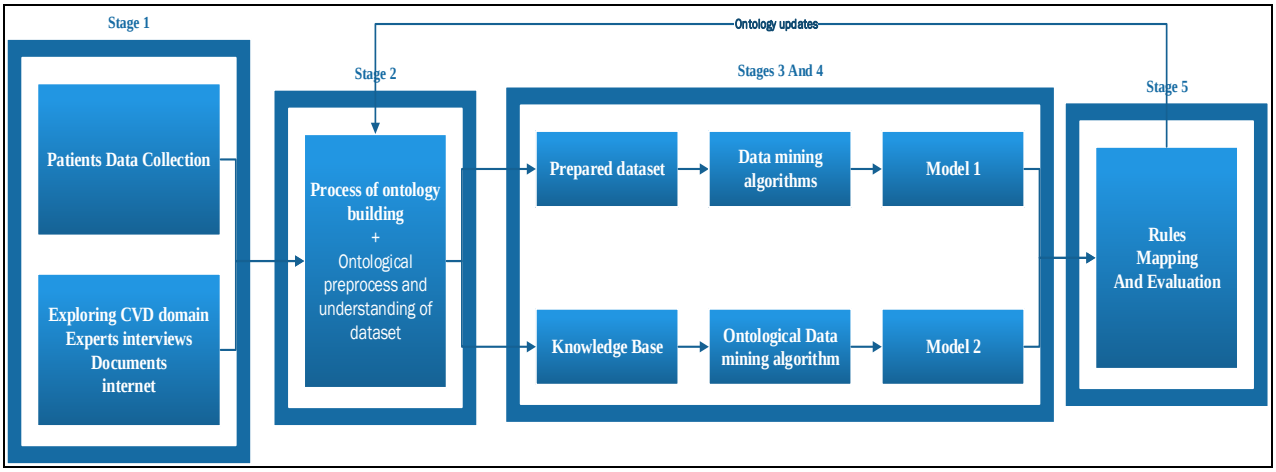


Figure 3.1 Methodology steps

3.2 Step 1: Collecting informations and Dataset:

The needed information are captured by analysis documents and papers related to the field and helding some interviews with cardio doctors who gave details about CVD and its symptoms. It's necessary to mention that, the collected information about cardiovascular disease is the information used by doctors in the Gaza Strip to diagnose the disease.

The data set collection was the heavy job, where Blood samples collected directly from patients from three different hospitals in Gaza strip during 2015-2016. Some needed data related to the patients are added to the data set from hospitals records. All the necessary tests was conducted for these samples at trusted and licensed lab in Gaza. The data set was collected for other issues and we can benefit from it in our work to help diabetes patient in avoiding CVD occurrence.

There are 40 columns (variables) and 300 rows (examples). The 40 variables can be sorted into three major categories. The first one related to personal information as shown in Table 3.1. The second related to Diabetes mellitus patient's diseases and treatments as shown in Table 3.2 .The third one is related to tests collected for research as shown in Table 3.3. It is about 20 tests including all normal tests for DM patients and other tests like Lp(a) and high sensitive (hs-CRP) were measured. Glucose, creatine kinase (CPK), aspartate aminotransferase (AST), lactate dehydrogenase (LDH), cholesterol, triacylglycerol's (T.G) and high density lipoprotein (HDL-C).

Table 3-1 Personal information attributes in dataset

NO	Attribute Name	Description
1.	Name	Name of the patient
2.	Address	The address of the patient
3.	Gender	Male or female
4.	Age	Age in years
5.	Education	Education level
6.	Smoking	Yes or no
7.	Exercises	Doses the patient practice exercises
8.	Type of disease	DM, DM-CVD ,RF ,CVD ,No

Table 3-2 Type 2 DM patient's diseases and treatments attributes

NO	Attribute Name	Description
1.	Diagnosis	DM, DM-CVD ,RF ,CVD ,No
2.	Years	When the patient has DM
3.	CVD	CAV types (heart failure, IHD, atherosclerosis, CVA)
4.	Thera	Sugar treatment (Insulin, Tablets, Diet, others)
5.	Lipdix	Lipdix therapy (yes or no)
6.	Test	Dose the patient make regular analysis
7.	Gtest	Dose the patient make glucose analysis
8.	Gtimes	Time to analysis glucose (monthly, 2 months ...)

9.	Lipid	Dose the patient make lipid analysis
10.	Ltimes	Time to analysis lipid (monthly, 2 months ...)
11.	Cadiac	cardiac tests
12.	Ctimes	Time to analysis cardiac (monthly, 2 months ...)

Table 3-3 laboratory tests attributes

NO	Attribute Name	Description
1.	FBG	The fasting blood glucose test
2.	Uric acid	Uric acid test
3.	TG	Triglycerides test
4.	creatinine	level of creatinine in the blood
5.	Total protein	total amount of albumin and globulin
6.	AST	aspartate aminotransferase, used to detect liver damage
7.	CK	creatine kinase, used to detect inflammation of muscles
8.	Albumin	is a protein made by the liver
9.	LPa	used by your doctor to measure the levels of Lp(a)
10.	CRP	high-sensitivity C-reactive protein
11.	Amylase	Enzyme produced by pancreas
12.	HDL3	Total blood (or serum) cholesterol
13.	ALP	Alkaline phosphatase enzyme
14.	Chol	Cholesterol test
15.	CK-MB	Creatine kinase-MB enzyme
16.	Electrolytes	Electrolytes are salts and minerals
17.	LDL	Low density lipoprotein
18.	MG	Magnesium
19.	LDH	Lactate Dehydrogenase, for an enzyme, or catalyst
20.	P-ANCA	Antibodies in white blood

3.3 Step 2: CVD Ontology building phase:

The general of semantic representations for knowledge discovery process in data mining depends on light weight ontologies. The light weight ontology are always superficial without solid relations among entities, but still easy to construct and facilitate the application in computer field greatly (Panov, Džeroski, & Soldatova, 2008). According to that we started to build a lightweight ontology from scratch, reusing other ontologies related to our domain.

There are several ways to develop and built ontology. METHONOLOGY methodology is considered one of the best and comprehensive ways to develop an ontology in ontology engineering (Al-Hamadani & Alwan, 2015), because it depicted design stages depth, which are interfere with each other and operates

simultaneously. Figure 3.2 shows METHONOLOGY, stages and development process (Al-Hamadani & Alwan, 2015).

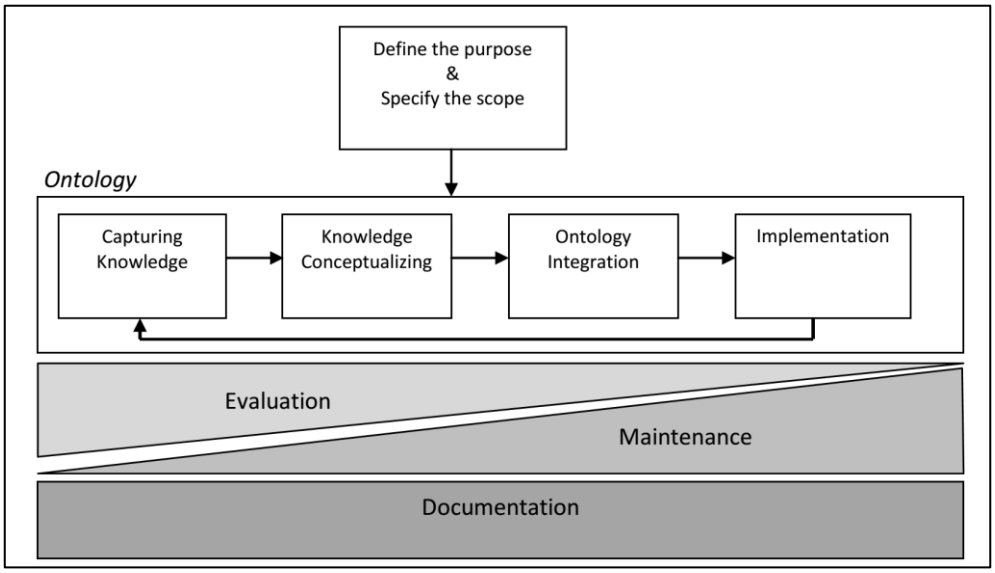


Figure 3.2 METHONOLOGY methodology(Al-Hamadani & Alwan, 2015)

Figure 3.2 shows the methodology for ontology development starts with defining the purpose of building ontology and specify the scope. The next step is knowledge capture and conceptualization. Then we start reusing known ontologies concern with our domain and implementing the generated ontology. Ontology evaluation and maintenance will be conducted to finalize the work of ontology building. The last step is documenting the constructed ontology to be reference and reusing it in the future.

3.3.1 Step 2.1: Ontology Construction framework:

Many formats that can be used to represent and construct ontologies has been developed in the recent years. These formats used for recording information possibility besides standardization aspects (Banaś, 2012). The most famous formats which based on XML such like :

- RDF (Resource Description Framework) – resource description language oriented to automatic processing؛
- OWL (Web Ontology Language) – a language being an extension of RDF, designed primarily for building ontologies؛

These languages have been designed primarily to create a so-called “Semantic Web”. And using these representation languages can build ontology directly, for which we can use even a simple programming, or XML editor. However using specialized and specific tools without need to focus on XML to create an ontology will be more appropriate, these tools protect the data from free manipulation. A good example an open platform is Protégé. Protégé framework interface show in Figure 3.3. Protégé a highly customizable package for ontology construction. It provides various tools such as graphical editors, classes, objects and entities which constitute the ontology. In the tools, you are more likely to find OWL looking more like a tree of classes and their descriptions. It also offers two popular formats of ontologies which are frames and OWL. To ensure the widest possible compatibility with other programs and technologies (Banaś, 2012).

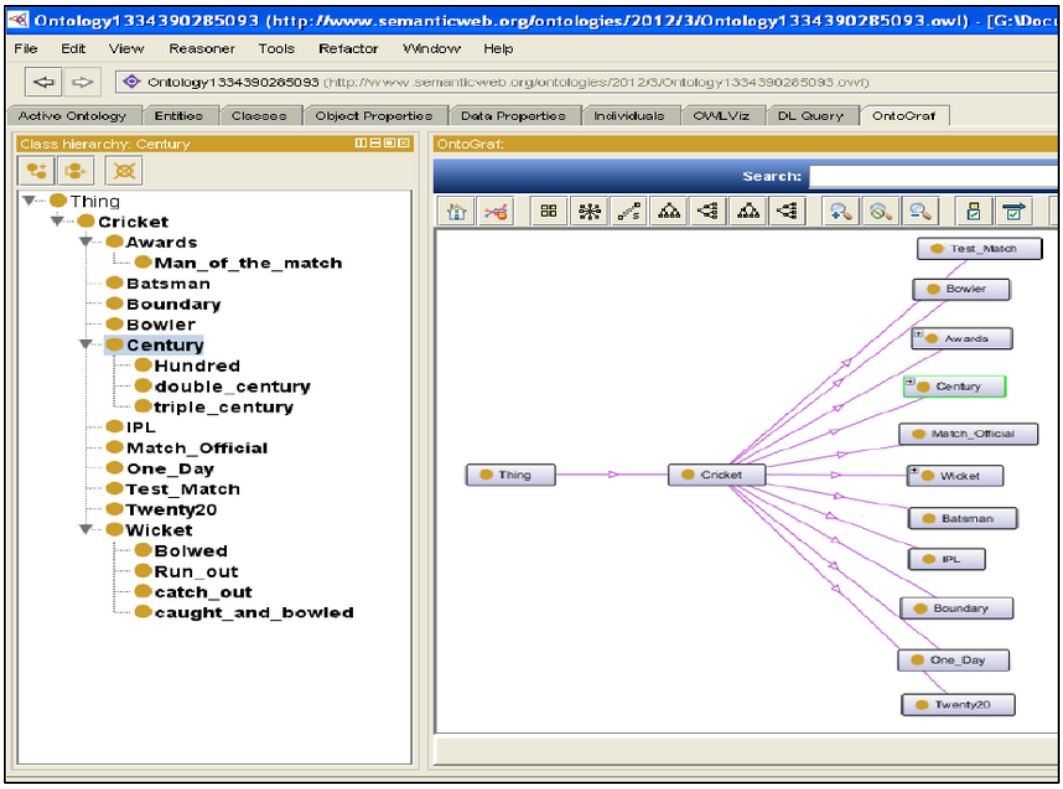


Figure 3.3 Protégé framework

3.3.2 Step 2.2: Ontology evaluation:

Evaluation methods will conducted to evaluate the proposed ontology. Expert evaluation will be conducted in addition to ontology reasoner (Pellet reasoner) embedded in portage environment for ontology design.

3.4 Step 3: Data mining phase

Data mining experiments conducted as in the following steps:

1. **Ontological Data understanding:** using ontology for the CVD domain is the main role for data understanding. Ontologies can help a new miner to be familiar with most important concepts and relationships, it also allow to recognize not obvious assumptions. For example we can benefit from mapping elements of an ontology on elements of the data scheme and vice

versa in identification of missing attributes that should be added to the data set and redundant attributes identification.

2. **Ontological Data preparation**: this phase is already connected with the previous Modelling phase. The dataset variables are categorized into semantic groups. Patient records are grouped into one record. Some unwanted attributes are removed according to ontology recommendations, and other are added.
3. **Data mining algorithms used are** :

Association rules: this algorithm used to express knowledge. Association rules are expressed in the form of IF-THEN rules.

Rule induction: which is one of the most important techniques of machine learning.

Decision tree: which is more powerful for classification problems and one of the predictive modelling approaches used in statistics, data mining and machine learning.

Model evaluation:

We will use different evaluation measure to evaluate the performance from different views. Accuracy, f-measure, and confidence will calculated for the association rules.

3.5 Step 4: Ontological Data mining phase:

Ontological Data mining algorithm found for mining sets of patterns for classification issues as proposed in (Ławrynowicz & Potoniec, 2014), called new data mining generation where patterns are represented as SPARQL queries over RDFS called sematic data mining. A data mining approach where domain ontologies are used as background knowledge.

The provider of the algorithm developed a tool that implements Fr-ONT-Qu algorithm, this tool called RMonto Ontological Extension to RapidMiner. This tool will be used in ontological data mining phase. The approach listed in Figure 3.4 explains the ontological data mining operation and how to mine ontological data set and generate model.

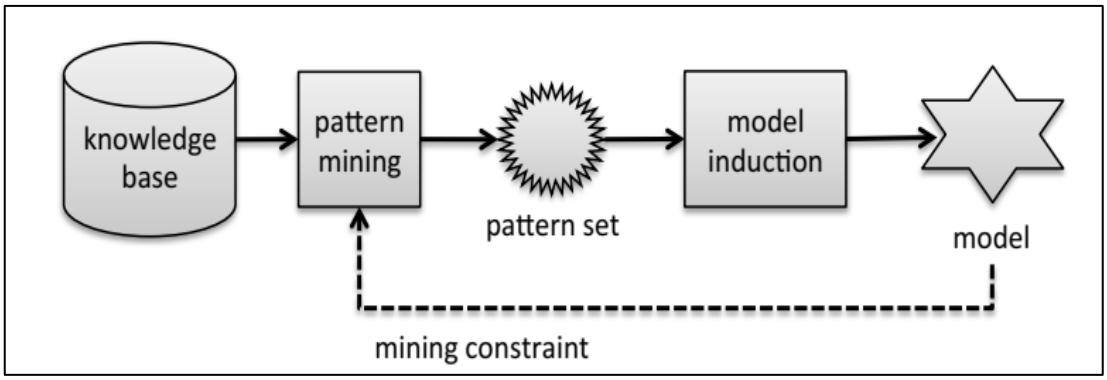


Figure 3.4 pattern-based classification method (Ławrynowicz & Potoniec, 2014)

We start with building CVD knowledge base and make it suitable for the pattern mining process. Then we specified the classes of ontology to be mined according to expert recommendation. Then SPARQL query used to select CVD patients. A pattern set generated, then we used the model induction method to generate the model for CVD patient. The generated model is constraint with the generated pattern mining from knowledge base.

3.6 Step 5: Models mapping and Evaluation phase:

This stage will be conducted by calculating the learning accuracy of the model from the rapid miner calculation and we will compare the result of this stage with data mining results from the previous stage to investigate about the correctness of the generated rules.

3.7 Summary:

The preceding chapter have presented the methodology of this research. This chapter is divided six sections, the first section represents the proposed methodology for discovering knowledge, and the second section shows the explanations of the data set. Then it shows how the methods of building ontology and the steps of the used methodology. Section four will explain the different data mining methods like association rules, rule induction and decision tree and how to evaluate the model. The last section illustrates ontological data mining algorithm and how to benefit from it in mining patterns.

Chapter 4

Cardiovascular Disease Ontology Implementation

Chapter 4
Cardiovascular Diseases Ontology Implementation

This chapter introduces the empirical and theoretical steps to build and develop specialized cardiovascular disease ontology from scratch, using expert knowledge, reused ontologies at the same domain and some medical documentations. This cardiovascular disease ontology will be used to improve the mining techniques based on the discovered pattern. In addition to that it will be used in ontological data mining approach. Ontology overview will be presented at section 1, section 2 will present ontology design steps, construction the ontology will be at section 3.

1.1 Overview

In knowledge engineering domain, ontology become the most important and promising topic, as it expresses the related basic domain concepts to the targeted field as well as it shows the relations in a consensual and understandable way (Cešpivová, Rauch, Svatek, Kejkula, & Tomeckova, 2004). It plays an important rules in knowledge discovery process (KDD). Before starting the mining process we start to build an ontology for cardiovascular disease. The variables used in this ontology collected from many medical resources, experts and reusing previous ontologies.

1.2 Classification of Cardiovascular disease

We started ontology building for cardiovascular disease from domain terms, which defined at Johns Hopkins Health System, Glossary - Cardiovascular Diseases(Medicine, 2018) , and asking some specialists about the domain and considering other reused ontology which will mentioned at section 4.7.

We have classified the first conceptualization terms of the cardiovascular disease into the following categories shown in Table 4.1, then this categories show to generate conceptualization tree.

Table 4-1 First conceptualization terms of the cardiovascular disease

Class Name	Description
Diseases	A disorder of structure or function in a human especially one that produces specific symptoms
Laboratory-Tests	A laboratory test checks a sample of blood, urine, other bodily fluid, or tissue to get more information about your health
Symptoms	Any subjective evidence of disease
Diagnosing	Tools are available to help you diagnose
Treatment	procedure that ends the medical condition
Patient-Info	Information about the targeted patients

1.3 Ontology classes classification and extraction

Each of these classes are has subclasses as follows:

1. Diseases Class

The diseases class refers to diseases types the affects the patient and heart function which includes:

1.1 CVD

1.1.1 Cardiovascular disease

1.1.2 Intradialytic hypotension (IDH)

1.1.3 Atherosclerosis

1.1.4 Cerebrovascular accident (CVA)

1.1.5 Heart failure

1.2 Diabetes Miletus

1.2.1 Type 2 Diabetes Miletus

2. Laboratory Tests

We have listed the most important laboratory tests as follow:

2.1.1 Lp (a)

2.1.2 high sensitive (hs-CRP)

2.1.3 Glucose

2.1.4 creatine kinase (CPK)

2.1.5 aspartate aminotransferase (AST)

2.1.6 lactate dehydrogenase (LDH)

2.1.7 Cholesterol

2.1.8 triacylglycerol (T.G)

2.1.9 High density lipoprotein (HDL-C).

3. Symptoms

Cardiovascular Disease patient general symptoms listed as follow:

3.1.1 Sudden chest pain

3.1.2 Effort chest pain

3.1.3 Shortness of breath

4. Diagnosing

Cardiovascular Disease patient diagnosing tools and procedures listed as follow:

4.1.1 Check heart rate

4.1.2 Check blood pressure

4.1.3 Physical exam

4.1.4 Check heart beat

5. Treatment

Cardiovascular Disease patient traditional treatments listed as follows:

5.1.1 Lipidex

5.1.2 Daily aspirin

5.1.3 ACE inhibitors

5.1.4 beta-blockers

6. patient information

Other personal factors for heart disease are listed below:

6.1.1 age

6.1.2 sex

6.1.3 smoking

6.1.4 family history

1.4 Defining terms and synonyms related to the topic Classification of Cardiovascular disease

keywords and their synonyms have been identified according to diagnosing of cardiovascular disease. These terms is a symptoms terms which is known to the most cardiologists. Some of these terms are: shortness of breath, chest paint, sudden chest pain, effort chest pain, high blood pressure.

At this stage we can't define the terms that are related to when the cardiovascular disease will happens These terms and synonyms will be pushed to the ontology after the mining process. The mining process will generates the causes of cardiovascular disease among diabetes patients and these generated rules will be pushed to the ontology for diagnosing issue. The design of ontology at this stage in order to be used in mining preparation and understanding process in addition to ontological data

mining process. There is now clear reasons that declare when the cardiovascular disease will occur, we will try to find them.

1.5 Building the hierarchy of concepts

Figure 4.1 shows domain concepts and relations extracted from a glossary definition for cardiovascular disease ontology which illustrates the core classes and relations between them.

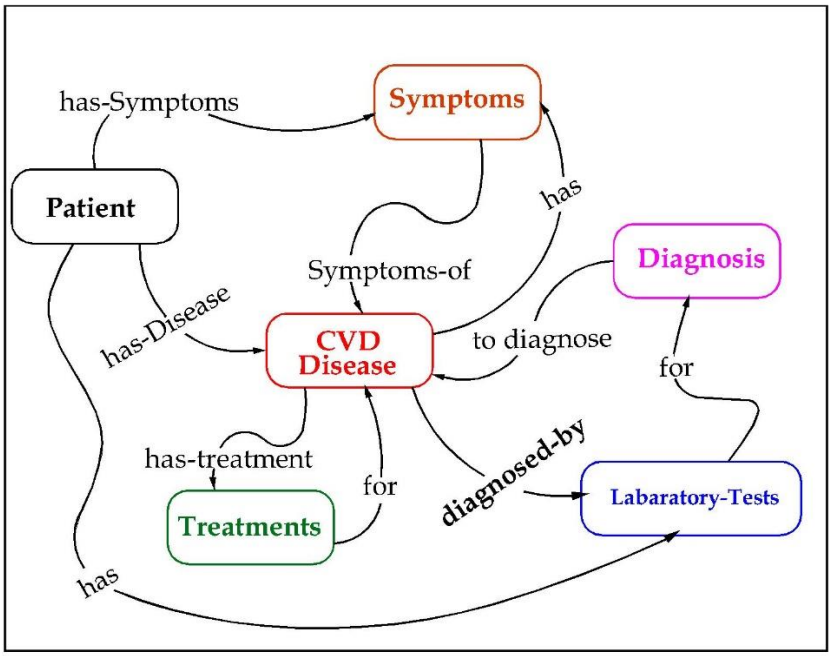


Figure 4.1 Ontology classes hierarchy

1.6 Relations between concepts and properties

The previous Figure 4.1 shows the relations between concepts of cardiovascular disease. The concept can have a set of sub concepts that can be contain more properties, some of these properties are illustrated in Table 4-2

Table 4-2 ontology concepts and properties

Relation name	Example
hasDiseases	Patient hasDiseases Cerebrovascular accident (CVA)
isSymptomOf	Sudden chest pain isSymptomOf Cerebrovascular accident (CVA)
hasSymptoms	Intradialytic hypotension (IDH) hasSymptoms Effort chest pain
hasSymptoms	Patient hasSymptoms Shortness of breath
treatmentFor	Lipidex treatmentFor beta-blockers

1.7 Ontology Construction

1. Purpose and scope stage is the starting point which specify the purpose of developing the ontology and its scope and will answer the following questions:
    1. What is the domain that the ontology will cover?
    2. For what we are going to use the ontology?
    3. For what types of questions should the ontology provide answers?
    4. Who will use and maintain the ontology?

2. We developed cardiovascular disease ontology to be knowledge representation for the domain and to prepare it for data mining tasks and ontological data mining, and the proposed ontology can be used by the specialist for diagnosing purposes of CVD based in reasoning process and inferences throw the ontology.

3. Knowledge capture stage to capture the knowledge required to build the ontology and choice of the candidate terms. At this stage, we analysis documents and papers related to the field and held some interviews with cardio doctors who gave details about CVD and its symptoms, diagnosing steps and treatments to define classes, objects and its properties, and the relations between them.

4. The conceptualization stage related to this previous stage, by producing the terms used on ontology development. These terms are helpful to build ontology classes and determining proprieties and relations between these classes.

5. The integration stage which responsible for benefit from previous known ontologies, and reusing when new ontology can be integrated to existing one. Updating some parts of an ontology, which designed to be updated. This does not change major concepts and structures but makes an ontology more precise. Our cardiovascular disease ontology reused some parts of existing ontology in (Al-Hamadani & Alwan, 2015), and other used in (Arsene, Dumitrache, & Mihu, 2011) and (Jović, Gamberger, & Krstačić, 2011) . These parts is treatments, symptoms and patient info classes and some of its instances. The reused parts has been updated to suite our lightweight ontology and the cardiovascular disease diagnosing procedure in Gaza strip.

1.7.1 Ontology classes in Protégé:

In section 4.2, we defined cardiovascular disease ontology classes and terms. We started the definitions of these main classes in Protégé as shown in Figure 4.2

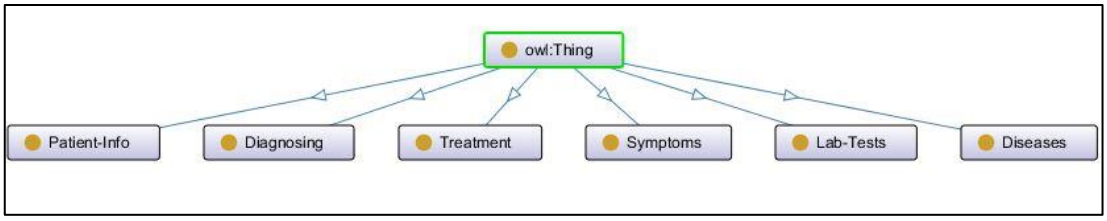


Figure 4.2 The Initial Class Hierarchy in protégé

1.7.2 Define properties of classes—slots

Classes alone will not provide enough information. Creating object properties plays important role in connecting classes (concepts) of the ontology. Properties define the internal structure of classes. Slots will likely be words which are not classes, we must assign each to a class (the most general one; remark that subclasses of a class inherit the slots) (Noy & McGuinness, 2001). Properties can be Intrinsic such as the patient has laboratory tests. It is also a relation between patient's class and laboratory tests class indicates that the patient must have laboratory tests for cardiovascular disease diagnosing. Figure 4.3 shows the graphical diagram for this relation.

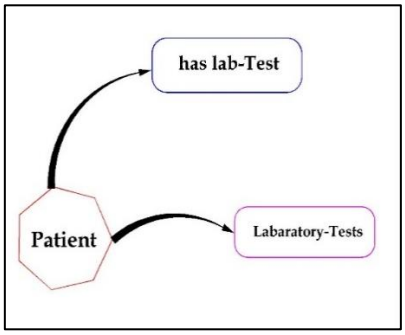


Figure 4-3 Example of a relation

Define properties of classes often interleaved with the previous step. The definitions of these main classes and subclasses in Protégé as shown in Figure 4.4 and Figure 4.5. The semantics of subClassOf demands that whenever A is a subclass of B, every property statement that holds for instances of B must also apply to instances of A. for example the main class named laboratory tests has property for diagnosing then any laboratory test will have this property. It makes sense to attach properties to the highest class in the hierarchy to which they apply. While attaching properties to classes, it makes sense to immediately provide statements about the domain and range of these properties. The defined object properties are illustrated in Table 4-3

Table 4-3 some of Objects property table

No.	Object property	Domain	Range
1	toDiagnose	Diagnosing	Diseases
2	diagnosedBy	Diseases	Diagnosing
3	treatmentFor	Treatment	Diseases
4	hasTreatment	Diseases	Treatment
5	hasDisease	Patient	Diseases
6	hasSymptom	Diseases	Symptoms
7	isSymptomOf	Symptoms	Diseases
8	hasLaboratoryTest	Patient	Laboratory Tests
9	forDiagonsing	Laboratory Test	Diagnosis
10	hasLowTestValue	Patient	Laboratory Tests
11	hasHighTestValue	Patient	Laboratory Tests
12	hasNormalTestValue	Patient	Laboratory Tests

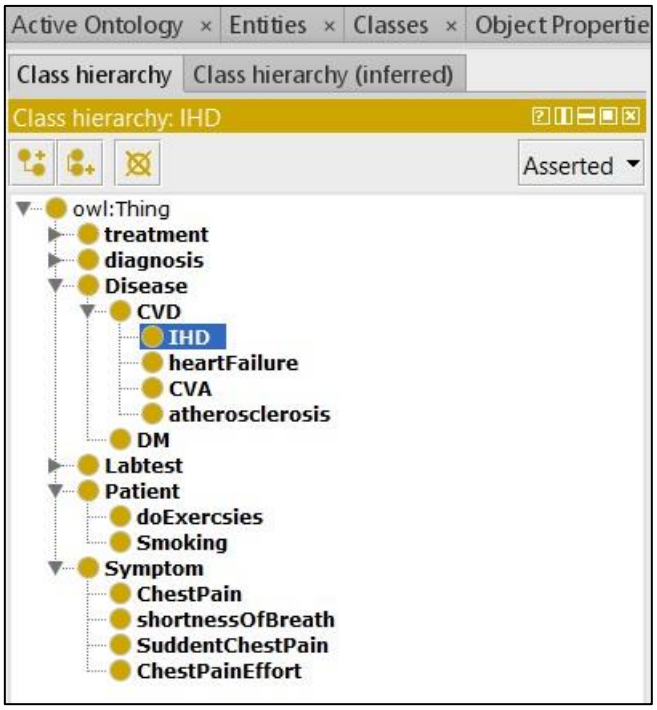


Figure 4-4 main classes and subclasses in Protégé

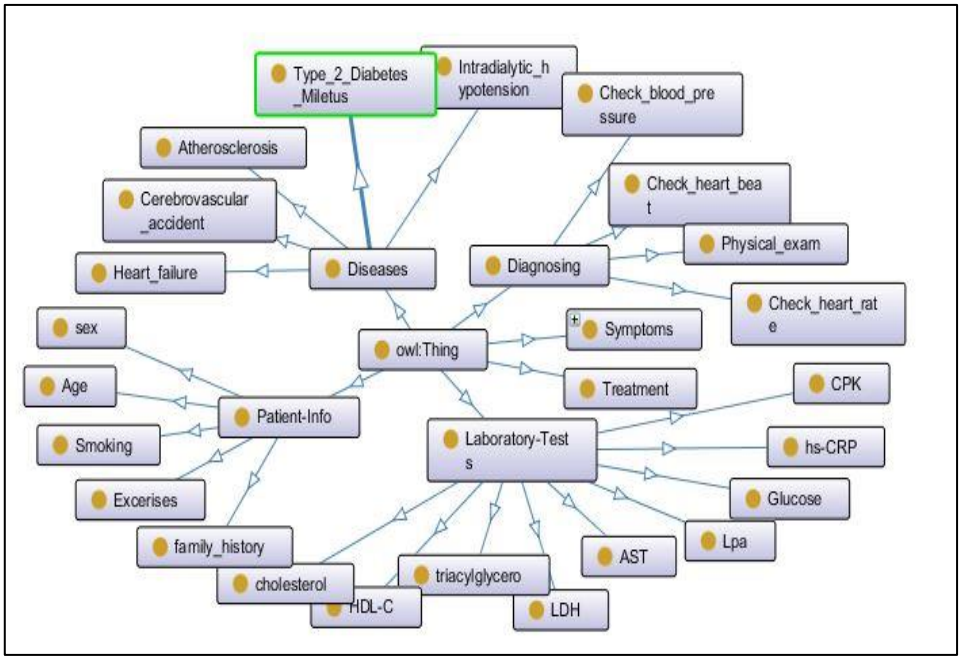


Figure 4-5 main classes and subclasses in Protégé

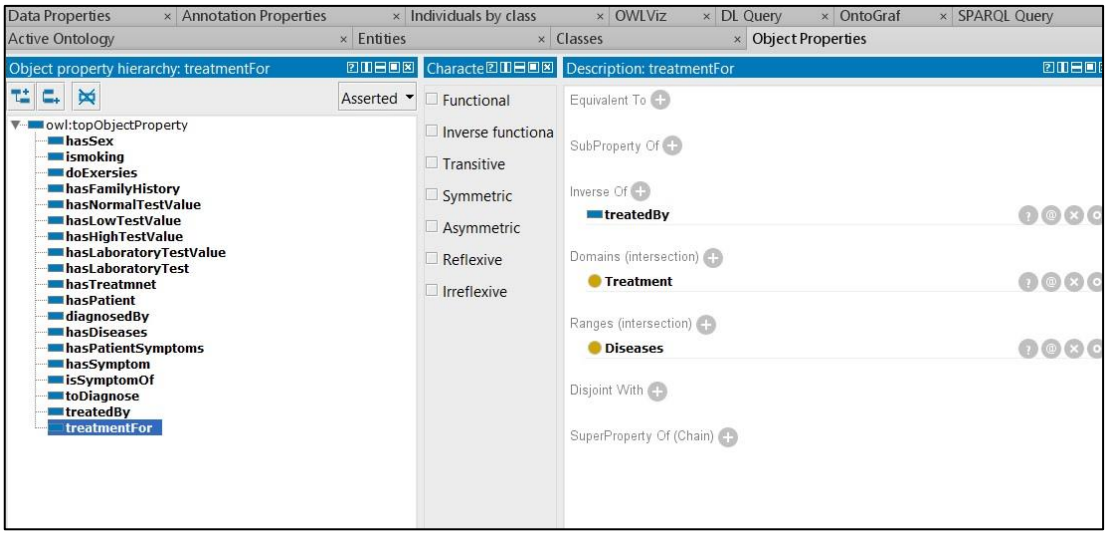


Figure 4-6 Defining objects property in Protégé

1.7.3 Define data properties:

Data properties describe the relationships between instances (individuals) and data values. Data properties such as is Smoking, hasName, DiseaseName and hasLp(a) are added to the ontology to link instances and classes and give values to instances in classes. Table 4.4 shows data properties of our cardiovascular disease ontology. Figure 4.7 shows show define data properties in protégé.

Table 4-4 data properties of cardiovascular disease ontology

No	Data property	Domain	Range
1.	AST	Patient	High, low, normal
2.	Chol	Patient	High, low, normal
3.	LP(a)	Patient	High, low, normal
4.	CK	Patient	High, low, normal
5.	CRP	Patient	High, low, normal
6.	FBG	Patient	High, low, normal
7.	HDL	Patient	High, low, normal
8.	LDH	Patient	High, low, normal
9.	TG	Patient	High, low, normal
10.	LDL	Patient	High, low, normal
11.	Name	Patient	String
12.	Sex	Patient	String
13.	DiseaseName	Patient	CDV&DM , CVD , DM
14.	isSmoking	Patient	Boolean
15.	has Age	Patient	Integer
16.	doExercises	Patient	Boolean
17.	familyHistory	Patient	Boolean
18	hasBloodPressure	Patient	Boolean

Slot-value type describes what types of values can fill in the slot. Here is a list of the more common value types: String, integer, Boolean, and Enumerated

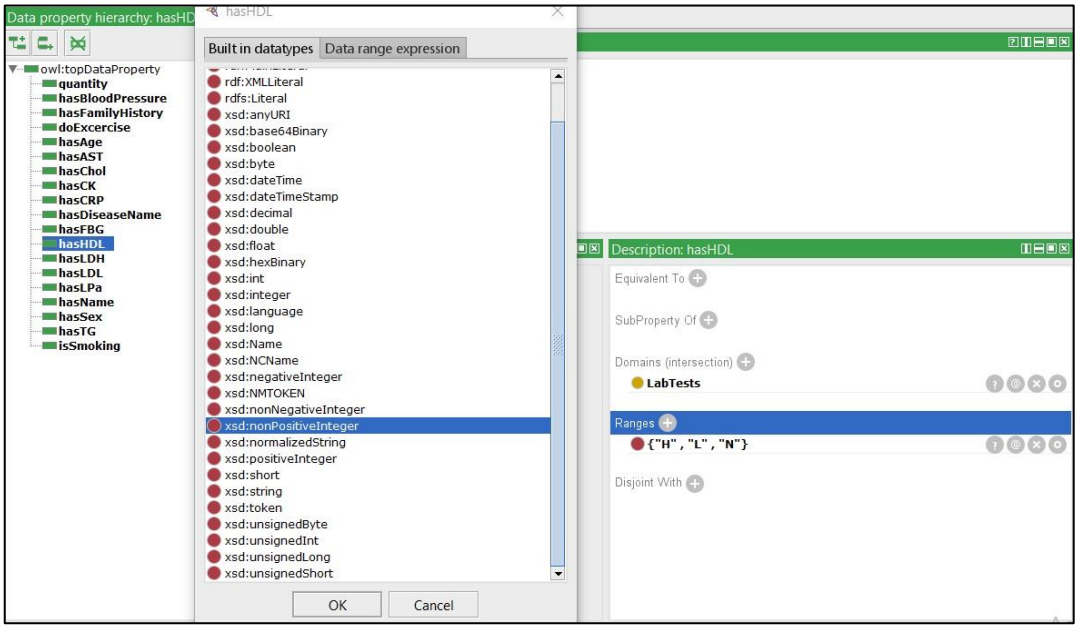


Figure 4.7 defining data Properties

1.7.4 Create instances

Instances are usually called Individuals in ontologies which represent objects in the domain. Individuals are specific instances of the concepts or objects. It represents the ground or atomic level of the ontology. The data and information of Individuals used in cardiovascular disease ontology are real data, collected from Gaza strip hospitals for CVD patients which illustrated at section 3.2. Figure 4.8 show the definitions of individuals in protégé. About 300 patient individuals was declared, and important laboratory set value added to the ontology individuals and other needed feature to be mined.

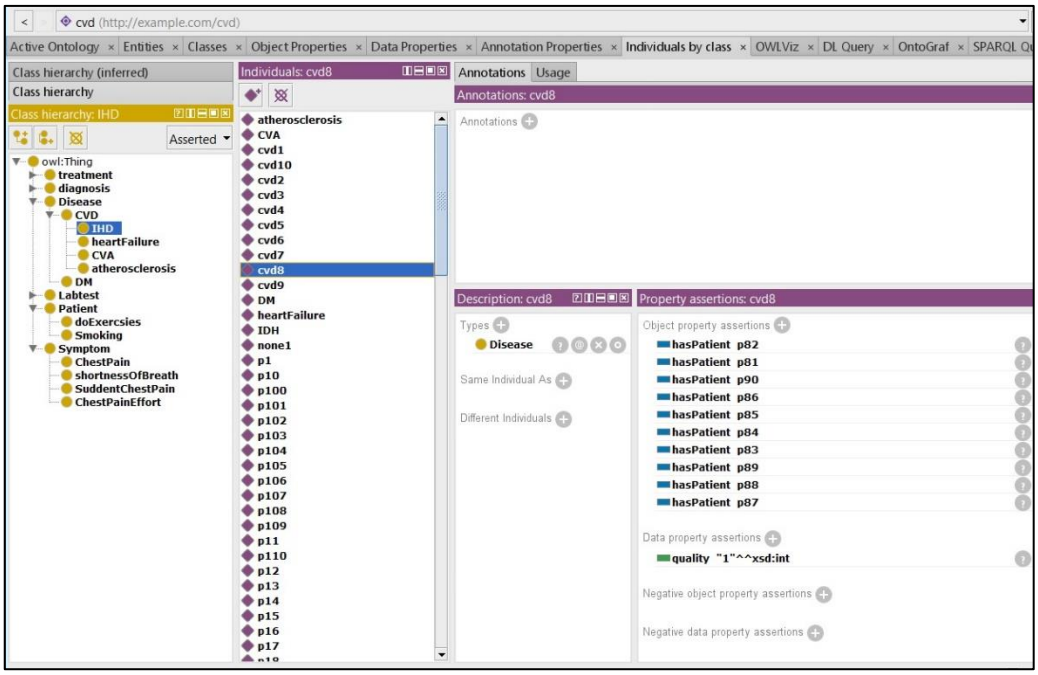


Figure 4.8 Definitions of individuals in protégé

1.7.5 Ontology Reasoner

Used to derive new facts from domain ontologies. Server reasoners are used last few years like: Pellet, Hermit, TrOWL ,Snorocket, RACER, CEL, ELK, SWRL-IQ, FACT++, and others (Abburu, 2012). It helps in inferring new rules and consequences, which are logical from asserted facts and axioms. It allow to find and produce different types of ontological relations such as symmetric, transitive, inverse and functional properties then use these new relations to add new fact to the domain and adding value to our ontology. We used the reasoner in this step is used to check the consistency of our ontology and to be used in ontological data mining at data mining stage later. In data mining experiments and results illustrated in chapter 5, we will use pellet reasoner in ontological data mining method, which needed in RMonto algorithm to mine knowledge base information.

1.8 Ontology Evaluation

In our case we depend on the domain expert to evaluate our proposed ontology. We asked him about the shortage in the ontology concepts/classes. The experts identified

the correctness of all classes. In addition to that we used ontology reasoner to check ontology consistency and stability and the results we satisfying.

1.9 Summary:

In this chapter we have shown how to develop a cardiovascular disease ontology using protégé software. The proposed ontology considered as a light weight ontology contains only the needed axioms and terms used in Gaza Strip for diagnosing CVD and used for our researching issues. The main goal of developing ontology is to be used in data mining process at the next chapters in addition to its role in sharing and reusing knowledge related to the CVD domain.

Chapter 5

Experimental Results

Chapter 5
Experimental Results

This chapter introduces the experimental results of data mining and ontological data mining experiments which are conducted on cardiovascular disease patient's dataset. The mining algorithms used are association rules method, rule induction method and decision tree assisted by ontology design to get higher accuracy learning models. Ontological data mining methods like pattern discovery method and rule induction will be conducted. We expected the usefulness of the result by comparing them with already known results, which known from a domain expert and data mining results and using different evaluation methods at the next chapter.

5.1 Experimental Settings:

The experimental environment that used for all the experiments was laptop with 4th generation core i7 CPU, 500GB HDD, with 8GB Ram.

5.2 Software and Tools

1. **Rapid Miner version 5.3**: Is an environment for experimenting with machine learning and data mining algorithms, where the experiments can be described visually as a process. Used for applying data mining methods and algorithms
2. **Microsoft Excel 2016:** Is a spreadsheet developed by Microsoft Company, for Windows OS, Mac OS X, and IOS. It features calculation, graphing, tables. Used for dealing with data set.
3. **Notepad++:** free source code editor, working on the Microsoft Windows environment. Notepad++ editor has Syntax Highlighting and Folding features. It used for opening OWL file and other text files.

5.3 Data Mining Experiments:

The design of the data mining experiment is concern with discovering rules, results, and predictions that concern to CVD. We expect the significance of the results by comparing them with the known associations derived from a domain expert. There are just two already known associations in chronic cardiovascular disease which are endorsed.

Type 2 Diabetes → Cardiovascular Disease

Smoking →Cardiovascular Disease

In this section the data mining experiment will be conducted, based on information provided by cardiovascular disease ontology, the results will be compared with known rules from experts, and evaluation methods will be conducted.

5.4 Association rules experiment:

Association rule methods started with applying FP-Growth algorithm. Figure 5.1 illustrates FP-Growth operator, which stands for frequency Pattern growth algorithm. This algorithm is workable on many kinds of data mining methods, and is necessary for association rule mining. Without having frequencies of attribute combinations, we cannot determine whether any of the patterns in the data often enough to occur and to be considered. Then association rules where generated.

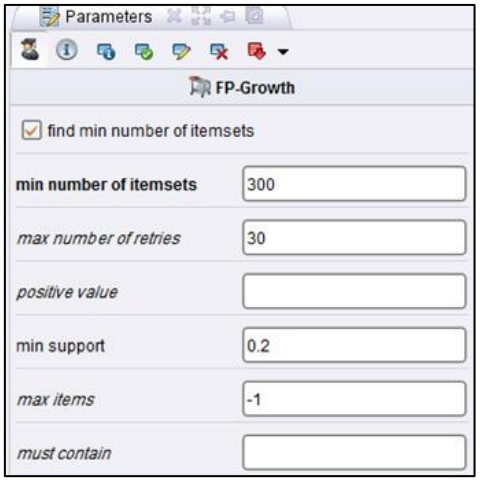


Figure 5.1 FP-Growth settings

The FP-Growth operator has some settings as shown in Figure 5.1. The min number of itemsets used in this experiment and the max number of tries parameter determines how many times the Operator may decrease the value before giving up. Increase this number to get more results. We decrease the min support value to increase the number of itemsets in the result.

Creating Association Rules operator in rapid miner can generate both a set of rules and a set of associated items, throw his output ports. Figure 5.2, shows the configuration of the association rules operator.

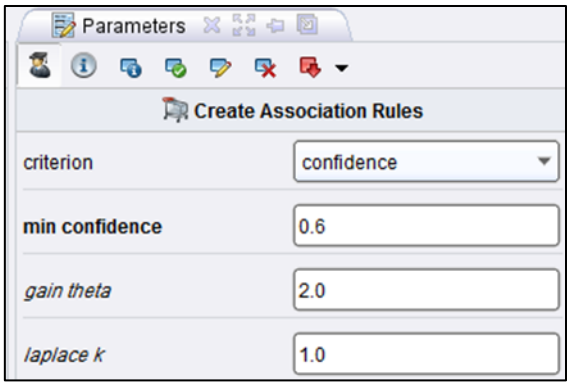


Figure 5.2 Association Rules operator settings

After running the experiment shown in Figure 5.3, which illustrates the overall process of the association rules mining operation we can find out the results in Figure 5.4.

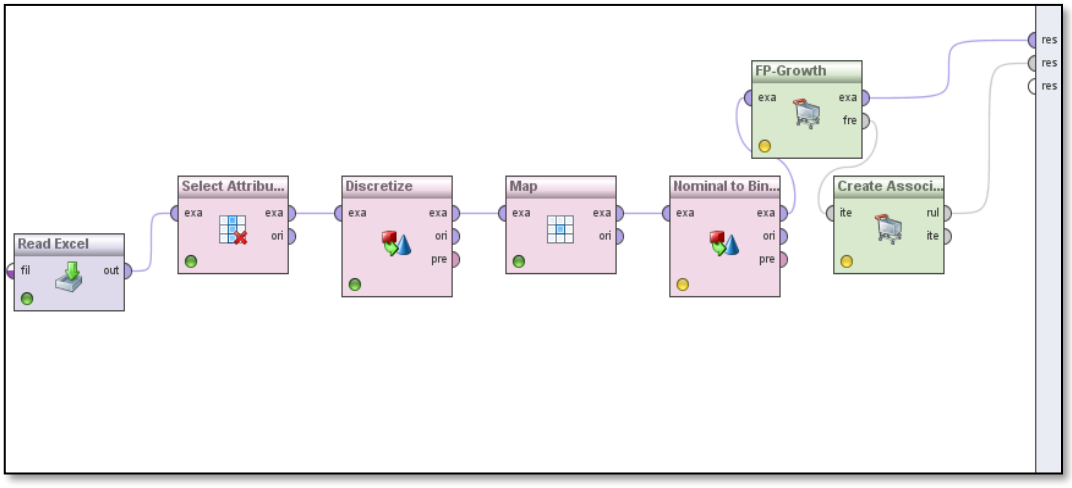


Figure 5.3 Overall association rule process

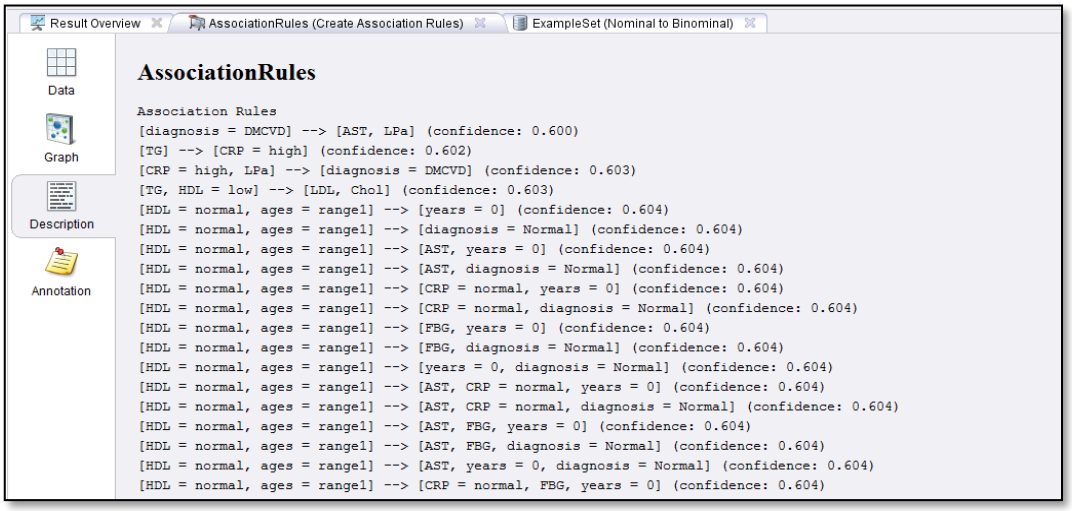


Figure 5.4 Association rules results

5.4.1 Association Rules Results:

We run the workflow with the following parameters:

Min support= 0.6 → parameter specifies the minimum support of the rules.

Confidence=0.6 → parameter specifies the minimum confidence of the rules.

Number of item set = 300 → parameter specifies the number of itemset.

Number of mined fields= 15 → parameter specifies the number of mined fields.

The number of generated rules = 2262→ the number of rules generated.

Best rules which dedicate CVD are:

Association Rule 1:

[LPa, diagnosis = DMCVD] --> [CRP = high] (confidence: 0.65)

[CRP = high, diagnosis = DMCVDL]--> [Lpa=high] (confidence: 0.677)

When the value CRP and LPA high, this means that, the patient with DM, will suffer from cardiovascular diseases in the future with confidence 67.7%

Association Rule 2:

[Diagnosis = DMCVD] --> [AST=high, LDH=high] (confidence: 0.660)

When the value AST and LDH, this means that, the patient with DM, will suffer from cardiovascular diseases in the future with confidence 66%

Association Rule 3:

[Diagnosis = DMCVD] --> [LPa, CK] (confidence: 0.650)

There is an association between LPa, CK, and cardiovascular diseases with confidence 65%

Association Rule 4:

TG--> [Chol = high] (confidence: 0.602)

There is an association between Triglycerides and Chol with confidence 60%

Association Rule 5:

[Diagnosis = DMCVD] --> [LPa, CRP = high] (confidence: 0.67)

There is an association between LPa and CRP with and CVD occurrence confidence 67%

The previous rules shows that there is a strong relation and association between LPa test = high, CRP test = high and diagnosis= Cardiovascular disease with confidence 67%

5.5 Decision tree experiment:

Decision tree is a classifier in the form of a tree structure looking to build tree with smallest levels and find the best attribute test to install at root, which results in greatest information gain with the average purity. Its represent rules, which can be understood by humans and used in knowledge system. It classify instances or examples by starting at the root of the tree and moving through it until a leaf node. The key requirements for decision tree algorithm are object should be expressible in terms of a fixed collection of properties or attributes and the target function has discrete output values and we should provide enough training cases to learn the model.

The first step in this experiment staring using rapid miner software by adding the (X-Validation) operator (see Figure 5.5) to estimate the accuracy of the experiment and estimate the error in practice.

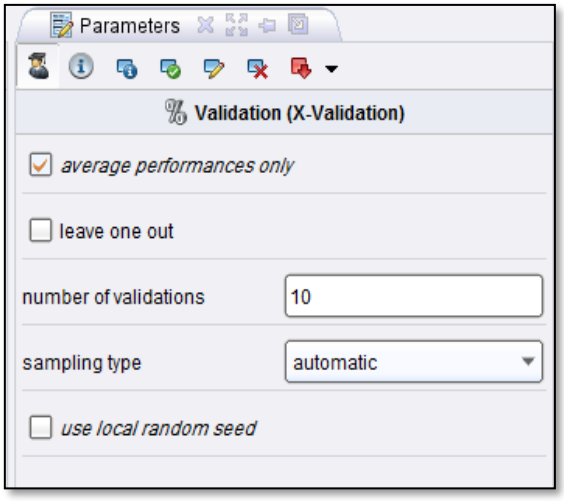


Figure 5.5 X-Validation operator settings

The next stage is a training stage starts by dragging decision Tree with following parameters:

Criterion = information gain

Minimal Size for split = 5

Minimal leaf size =2

Minimal gain= 0.15

Max-depth= 6

Confidence= 0.5

Pruning enabled

The information gain is to measures how well a given attribute separates the training examples according to their target classification. And this measure is used to select among the candidate attributes at each step while the tree are growing. Information gain measures the expected reduction in entropy, or uncertainty.

The last stage is testing stage. It starts by dragging the apply model operator. Which takes an Example Set as input and gives a Performance Vector as output .The overall process of the decision tree experiment shown in Figure 5.6

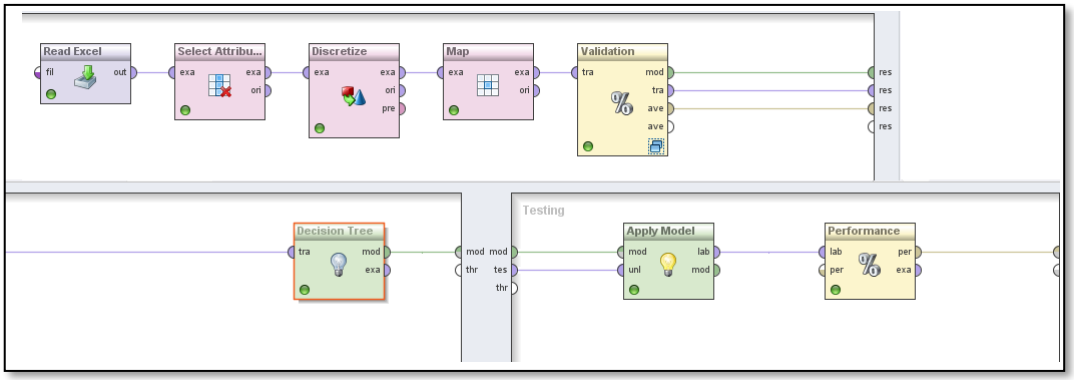


Figure 5.6 Overall process of the decision tree

5.5.1 Decision tree generated model:

When we applied decision tree method on our corpus, we got the tree in Figure 5.7

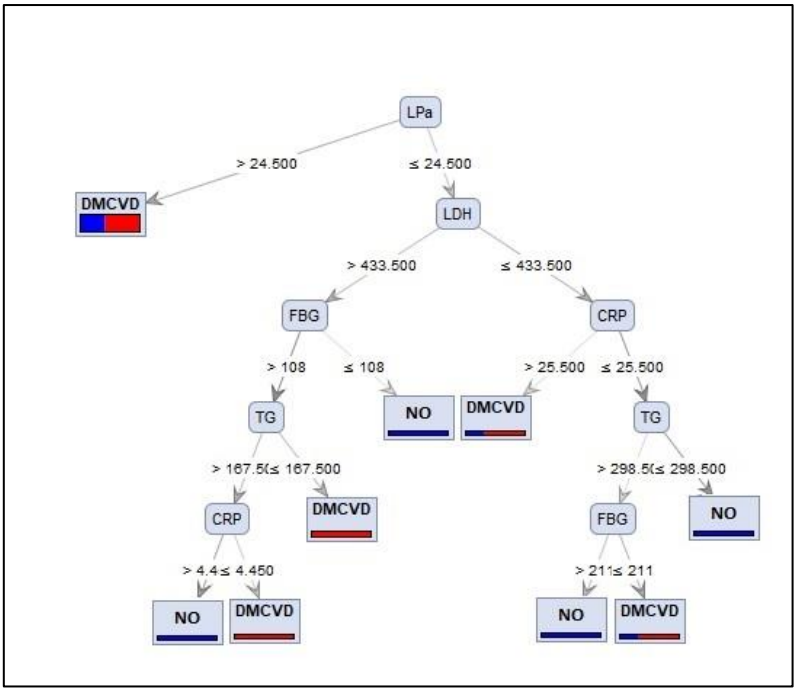


Figure 5.7 Decision tree algorithm output

Figure 5.7 shows the generated tree which conclude that LPa laboratory test is the most important factor to detect CVD. When it has high value the diabetes patients will be ready to be cardiovascular disease patient strongly.

5.6 Rule Induction method

Rule sets, induced from data sets, are used mostly to classify new, unseen cases. We used rule induction method in rapid miner on cardiovascular disease patient dataset to get the rules as follow:

The overall process of rule induction at rapid miner shown at Figure 5.8 and the process run with the following parameters:

Criterion = information gain

Number of validation = 30

Sampling type = shuffled

Pureness = 0.7

Minimal prune benefit = 0.2

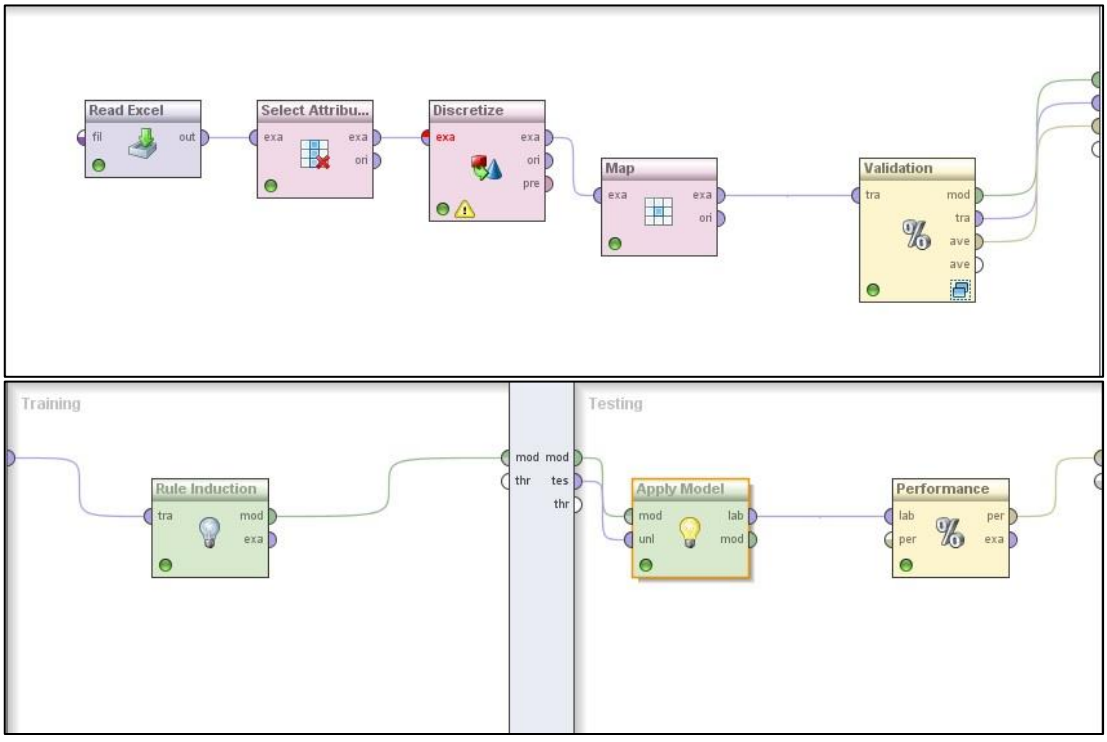


Figure 5.8 Overall process of rule induction

The results of Rule Induction method:

RuleModel

If	LPa > 41	and	FBG >= 131	then	DMCVD
If	LDH > 442	and	CK > 136	then	DMCVD
If	TG <= 146	and	HDL > 36	then	DMCVD

5.7 Ontological data mining approach

In this section, we will explain the steps of using ontological data mining algorithm and rapid miner plugins to mine the cardiovascular disease ontology, which explained at chapter 4 and using the data set explained at chapter 3.

5.7.1 Tool preparations:

Before we starts the mining process, we have to prepare the work environment by importing the needed tools. RMonto, which indicates to "Ontological Extension to RapidMiner", is the used tool and imported to rapid miner environment as shown in Figure 5.9. The tool includes ontology as root of the tree, load file used to load OWL file, which is the data set to be mined. It includes also a build knowledge base operator and Fr-ONT-Qu algorithm in pattern mining folder in the imported plugins.

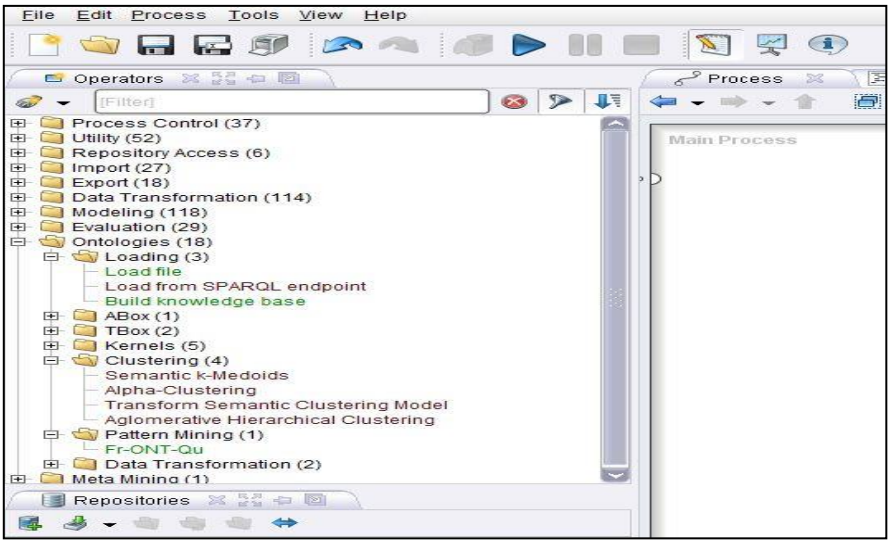


Figure 5.9 RMonto Ontological Extension plugins

5.7.2 Pattern discovery method

We started the workflow by creating ontology individuals in protégé software and adding all patient information including personal information, disease type, symptoms, and laboratory tests results for each patient and other needed information. This method shown in workflow at Figure 5.10 will explain how we build complex features by relational pattern mining from ontologies. We start the workflow at Figure 5.10 with adding Load file to load CVD owl file (Ontology from protégé), then we added the Build knowledge base operator and connecting the first one to the second one then adding Fr-ONT-Qu operator, which is the pattern-mining operator. Then connect Fr-ONT-Qu operator to knowledge base input to the Build knowledge base output and providing the knowledge base with knowledge about the patients by providing Fr-ONT-Qu operator with classes the represents properties of the patient. The SPARQL operator added to select the learning examples. The Propositionalisation operator transforms learning examples and frequent pattern into attribute-value table. The target of this experiment (Frequent Pattern Discovery) is to discover pattern about patients with respect to cardiovascular disease using the abstract rule has_disease. CVD ontology main classes are used to define the proposed ontology to the model.

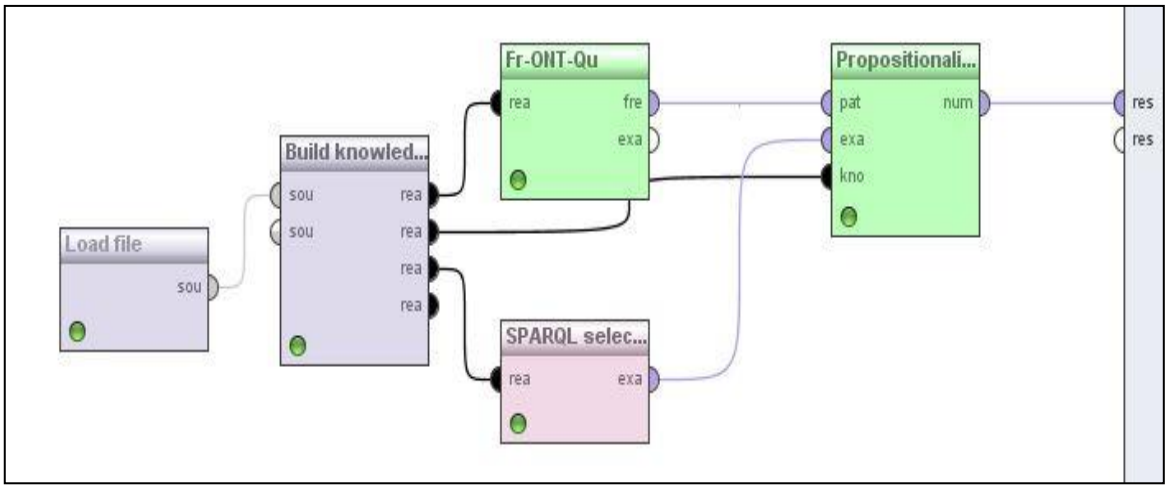
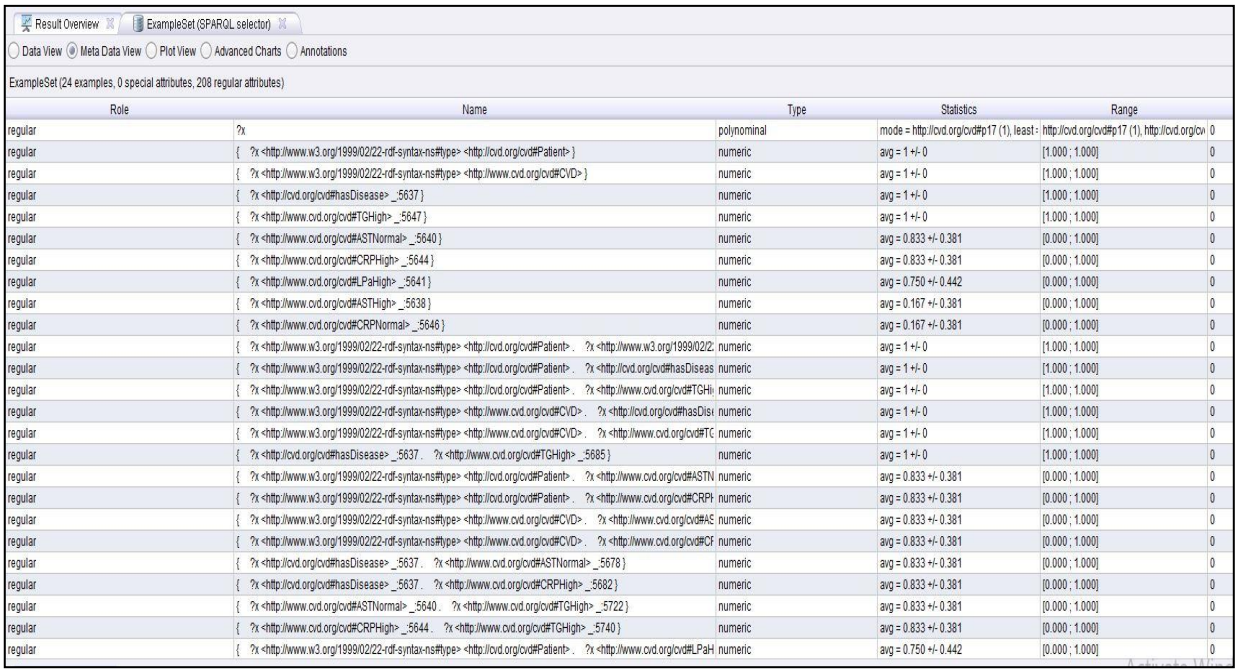


Figure 5.10 Pattern discovery workflow

5.7.2.1 Pattern Discovery Results

After running the workflow, the results characterized and explained by the corresponding selection of concepts in the domain ontology as shown in Figure 5.11 and explained below. The patients who have disease (Cardiovascular disease) in the data set are listed at the results and the attributes for every patient including its laboratory tests results stored at the ontology but in the form of graph pattern.



Result Overview | ExampleSet (SPARQL selector)

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (24 examples, 0 special attributes, 208 regular attributes)

Role	Name	Type	Statistics	Range	
regular	?x	polynominal	mode = http://cvd.org/cvd#p17 (1), least =	http://cvd.org/cvd#p17 (1), http://cvd.org/cv	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://cvd.org/cvd#Patient> }	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.cvd.org/cvd#CVD> }	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://cvd.org/cvd#hasDisease> _:5637 }	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://www.cvd.org/cvd#TGHigh> _:5647 }	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://www.cvd.org/cvd#ASTNormal> _:5640 }	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.cvd.org/cvd#CRPHigh> _:5644 }	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.cvd.org/cvd#LPaHigh> _:5641 }	numeric	avg = 0.750 +/- 0.442	[0.000 ; 1.000]	0
regular	{ ?x <http://www.cvd.org/cvd#ASTHigh> _:5638 }	numeric	avg = 0.167 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.cvd.org/cvd#CRPNormal> _:5646 }	numeric	avg = 0.167 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://cvd.org/cvd#Patient> . ?x <http://www.w3.org/1999/02/2	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://cvd.org/cvd#Patient> . ?x <http://cvd.org/cvd#hasDiseas	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://cvd.org/cvd#Patient> . ?x <http://www.cvd.org/cvd#TGHi	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.cvd.org/cvd#CVD> . ?x <http://cvd.org/cvd#hasDis	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.cvd.org/cvd#CVD> . ?x <http://www.cvd.org/cvd#TG	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://cvd.org/cvd#hasDisease> _:5637 . ?x <http://www.cvd.org/cvd#TGHigh> _:5685 }	numeric	avg = 1 +/- 0	[1.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://cvd.org/cvd#Patient> . ?x <http://www.cvd.org/cvd#ASTN	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://cvd.org/cvd#Patient> . ?x <http://www.cvd.org/cvd#CRPH	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.cvd.org/cvd#CVD> . ?x <http://www.cvd.org/cvd#AS	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.cvd.org/cvd#CVD> . ?x <http://www.cvd.org/cvd#CF	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://cvd.org/cvd#hasDisease> _:5637 . ?x <http://www.cvd.org/cvd#ASTNormal> _:5678 }	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://cvd.org/cvd#hasDisease> _:5637 . ?x <http://www.cvd.org/cvd#CRPHigh> _:5682 }	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.cvd.org/cvd#ASTNormal> _:5640 . ?x <http://www.cvd.org/cvd#TGHigh> _:5722 }	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.cvd.org/cvd#CRPHigh> _:5644 . ?x <http://www.cvd.org/cvd#TGHigh> _:5740 }	numeric	avg = 0.833 +/- 0.381	[0.000 ; 1.000]	0
regular	{ ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://cvd.org/cvd#Patient> . ?x <http://www.cvd.org/cvd#LPaH	numeric	avg = 0.750 +/- 0.442	[0.000 ; 1.000]	0

Figure 5.11 Pattern discovery results

Some of the results from rapid miner shows that:

1. {?x <http://example.com/cvd#hasPatient> _:51461 . _:51461 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#LPaHigh> . _:51461 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#TGHigh> }

2. {?x <http://example.com/cvd#hasPatient> _:51461 . _:51461 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#LDHHigh> . _:51461 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#TGHigh> }

3. { ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#Disease> . ?x <http://example.com/cvd#hasPatient> _:48891 . _:48891 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#FBGHigh> . _:48891 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#Patient> }

4. { ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#Disease> . ?x <http://example.com/cvd#hasPatient> _:48891 . _:48891

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://example.com/cvd#CRPHigh> . _:48891
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://example.com/cvd#Patient> }

5. { ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://example.com/cvd#Disease> . ?x <http://example.com/cvd#hasPatient>
_:48891 . _:48891 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://example.com/cvd#LPaHigh> . _:48891 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#Patient> }

5.7.3 Rule induction method

The workflow on discovering rules from ontological data set will be applied on a dataset of laboratory tests for cardiovascular diseases patients, which illustrated in chapter 3 and explained at table 3.4. We started the workflow as in section 4.3 by creating ontology individuals in protégé software and adding all patient information. The next step starting at rapid miner software by importing the owl file and adding build knowledge base operator and connect it to Fr-ONT-Qu operator to the Build knowledge base output. To discover patterns about patient with respect to laboratory tests, symptoms, disease, diagnosing etc. All classes URIs are defined in the class parameter. Then add SPARQL selector with the following query:

select distinct ?x
where {
?x <http://example.com/cvd#has_patient>?y
}

The Propositionalisation operator, which transforms learning examples and frequent pattern into attribute-value table. Finally the rule induction operator to generate the rules. The overall process of the rule induction work flow shown in Figure 5.12

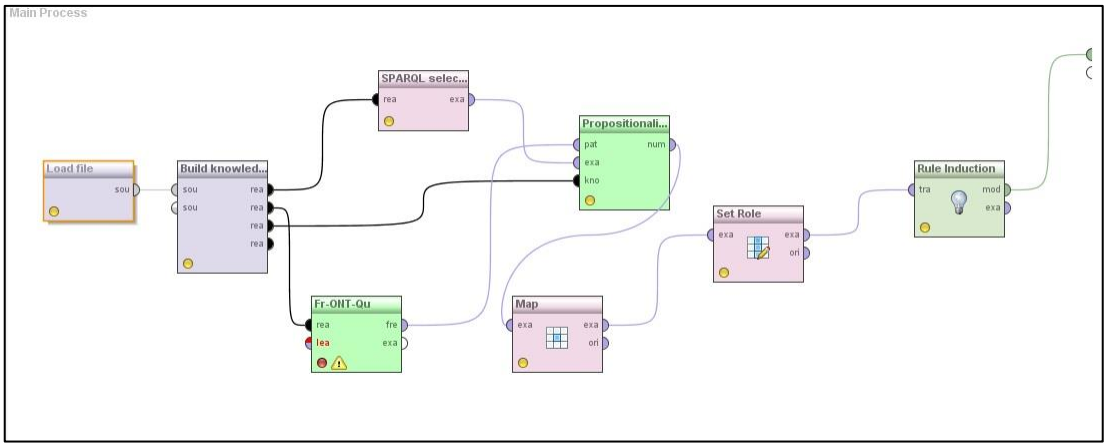


Figure 5.12 rule induction workflow

5.7.3.1 Rule induction results

After running, the workflow explained in Figure 5.12, the results characterized and explained by the corresponding selection of concepts in the domain ontology as shown in Figure 5.13. The figure shows that there are about 273 rules generated from this experiment. These rules reflects the effectiveness of some laboratory tests on all diabetes patients to be a cardiovascular disease patient and put their life at risk.

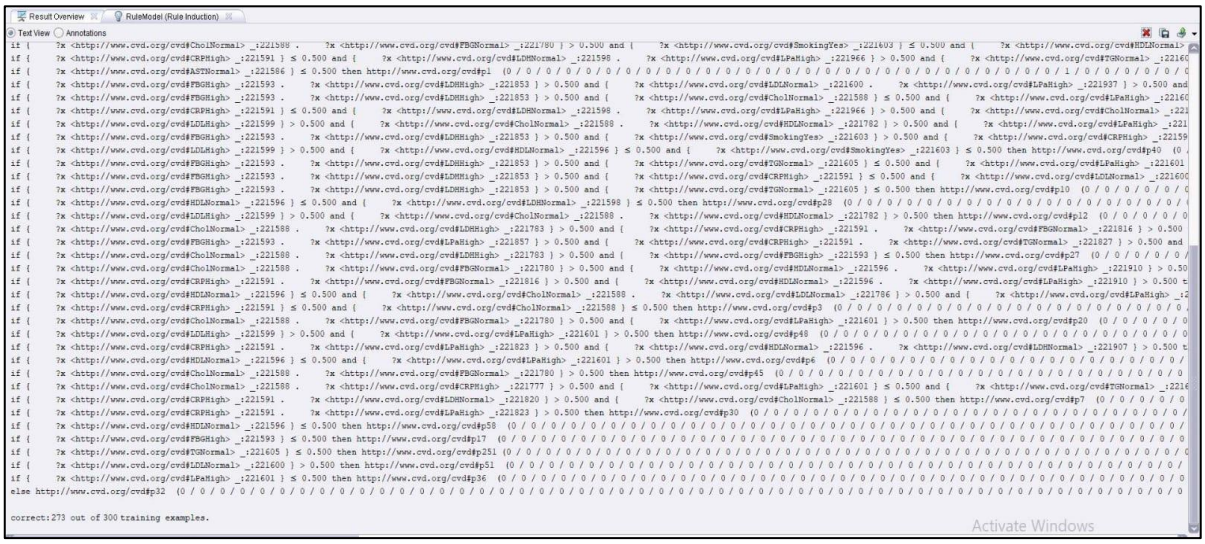


Figure 5.13 Results of process running

We can note from Figure 5.13, that the system generated many rules, which is difficult to analyze and understand. Therefore, we divided patients into 10 groups within the ontology in order to reduce the number of rules and ignore unneeded ones. The division step will enhance the model and give rules that are more accurate. The number of rules reduced to be five rules after this step as shown below:

RuleModel:

.1 if { ?x <http://example.com/cvd#hasPatient> _:51163 . _:51163 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#FBGHigh> . _:51163 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#LPaHigh> } > 0.500 then 1

.2 if { ?x <http://example.com/cvd#hasPatient> _:51163 . _:51163 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#HDLHigh> } > 0.500 then 1

.3 if { ?x <http://example.com/cvd#hasPatient> _:51163 . _:51163 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#CRPHigh> . _:51163 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#LPaHigh> . _:51163 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#TGHigh> } > 0.500 then 1

.4 if { ?x <http://example.com/cvd#hasPatient> _:51163 . _:51163 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.com/cvd#TG> } > 0.500 then 1 else http://example.com/cvd#normal then 0

Correct: 10 out of 10 training examples.

6.1.1 Summary

The preceding chapter have presented the experimental methods for data mining and ontological data mining algorithms. We used rapid miner software to implements all the experiments and added needed plugins.

The next chapter will analysis the results and shows the recommendations.

Chapter 6

Evaluation and Recommendation

Chapter 6

Evaluation and Recommendations

6.1 Association Rules Results Analysis and evaluation:

Using ontology along with the data mining operation, helped to skip some hypotheses, which would not make any sense from ontological point of view. In addition to that the use of ontology at the pre-mining phase has minimized the sum of mined attributes in the data set, thus the number of generated rules will be decrease. In our experiment, we noticed that it has decreased to 60% with the same setting to the experiment. Since we mined the overall data without any minimizing to compare the rules sum.

From the results show in Figure 5.4 and explained in section 5.4, we notice that patients with Diabetes Susceptible to CVD disease affected by high LPa and CRP. Also, by high LDH and AST, high Chol and TG, and there is an association between Lpa and CK. The confidence of each rule are explained at Table 6.1 and Figure 6.1

Table 6-1 Association Rules Confidence

number	Affected Laboratory Test	Confidence
Rule 1	LPa and CRP=high	65.0%
Rule 2	CRP=high and LPa=high	67.7%
Rule 3	AST=high and LDH=high	66.0%
Rule 4	LPa and CK	65.0%
Rule 5	TG, chol=high	60.2%

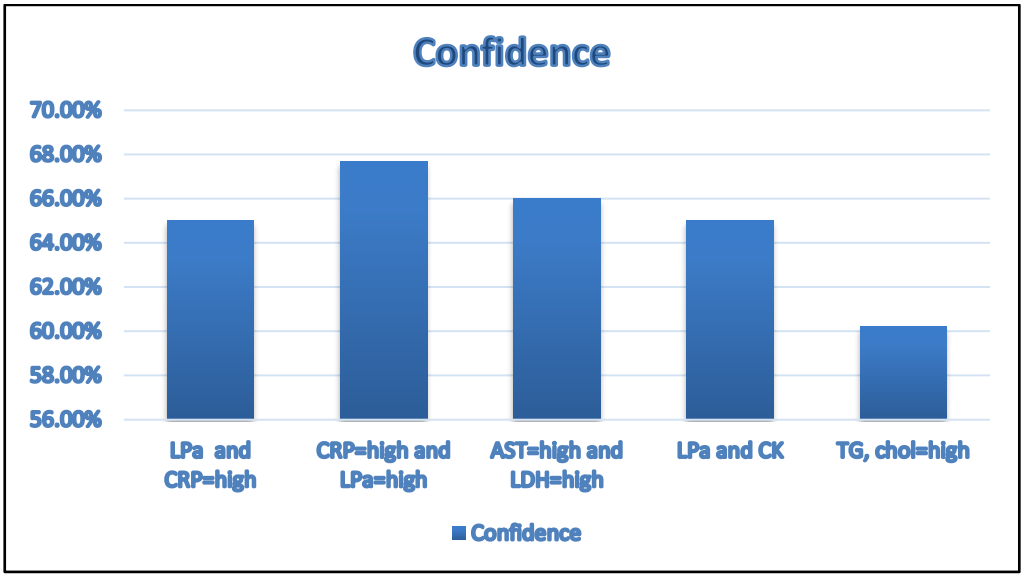


Figure 6.1 Association Rules Confidence

6.2 Decision tree Results Analysis and evaluation:

From the generated tree shown in Figure 5.7 , we conclude that LPa laboratory test is the most important factor to detect CVD, when it has a value = high it most probable the patient may got CVD. If the value of LPa laboratory test is normal, the factor that can detect CVD could be LDH laboratory test if it is =high with FBG and TG laboratory tests = high. If LDH is not the factor then we can look at CRP laboratory test if it is = high, then it may get CVD. Finally, if not CRP we can look at TG laboratory test if it is high then the patient may get CVD.

The results above is accurate by 88.25%

F-measure is 88.77%

The confusion matrix shown in Table 6.2 and the accuracy of tanning explained in Figure 6.1

Table 6-2 Decision Tree confusion matrix

	True CVD=NO	**True CVD =YES**
Predicted CVD=NO	169	16
Predicated CVD= yes	31	184

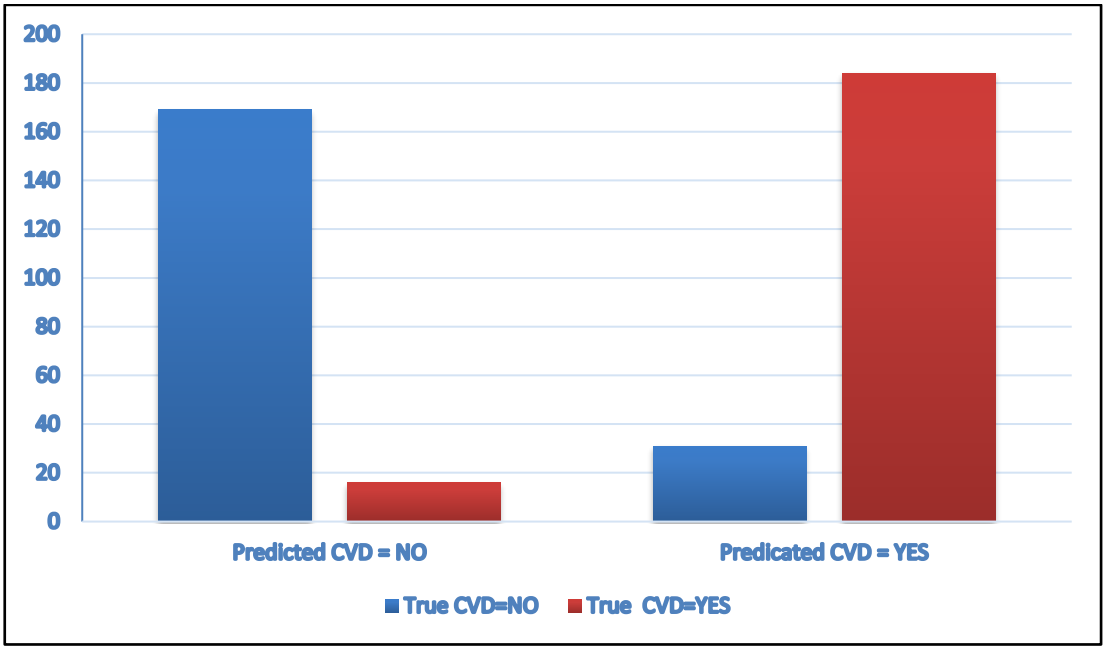


Figure 6.2 Decision Tree Accuracy

As shown in Table 6.2 and Figure 6.2 we can conclude that the model success to predict 184 patient of CVD correctly and not able to classify 16 patient.

6.3 Rule Induction Results Analysis and evaluation:

From the rules generated and shown at section 5.6, we can conclude that LPa with value greater than 41, which means high value, combined with FBG with value more than 131 which also high value, with get CVD. Also, if LDH greater than 442 which considered as a high laboratory test value and CK with more than 136 which considered as a high laboratory test value, this means that the patient will have CVD. In addition, if TG less than 146 and HDL greater than 36 this means that the patient will have CVD.

These rules is accurate with 80.75%

F-measure equals 83.84%

The confusion matrix show at table stated in Table 6.3 and explained at Figure 6.3

Table 6-3 Induction Rules Accuracy

	True CVD=NO	True CVD =YES
Predicted CVD=NO	137	14
Predicated CVD= yes	63	186

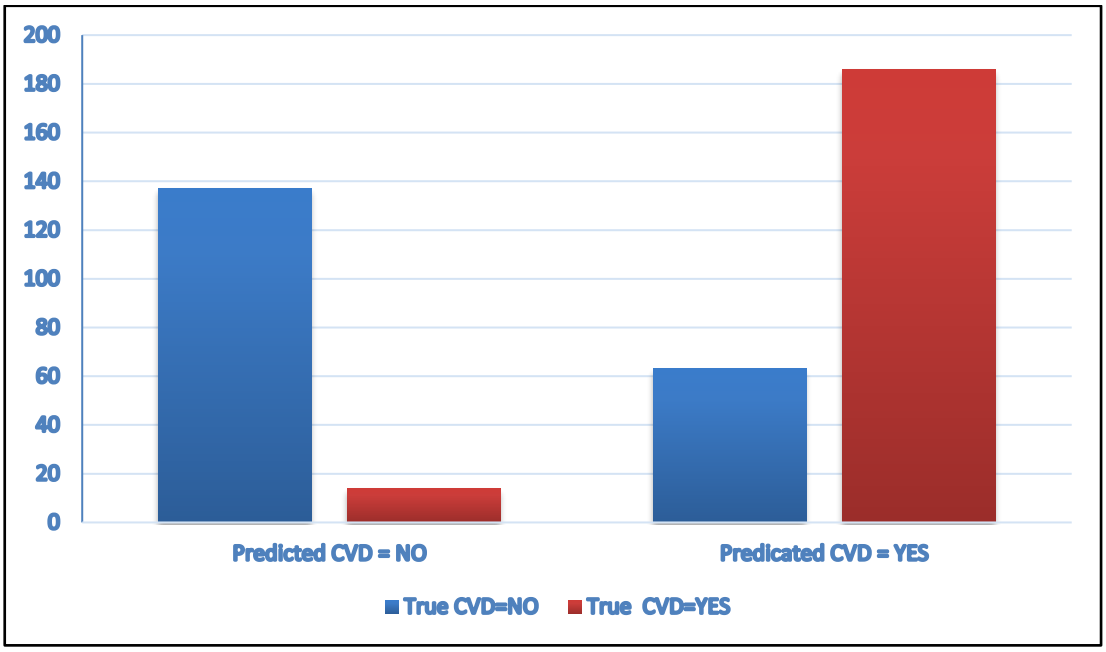


Figure 6.3 Rule Induction Accuracy

As shown in Table 6.3 and Figure 6.3 we can conclude that the model success to predict 186 patient of CVD correctly and not able to classify 14 patient.

6.4 Ontological Pattern Discovery Results Analysis and evaluation:

In this experiment, we have mined a set of pattern by mining knowledge encoded in domain ontologies rather than only purely empirical data. This method enables working directly on structured, relational data and effects on the results of data mining. Additionally, its custom algorithm implementations combined with the power of Rapid Miner through transformation/extraction from the ontological data to attribute-value data.

The results of experiment show that the high results of LPa, CRP, LDH, and TG are the most important factors on cardiovascular disease occurrence.

6.5 Ontological Rule induction Results Analysis and evaluation:

From the rules generated at section 5.7.3 we can see that diabetes mellitus patient will be Strong candidate to have a cardiovascular disease, or to have CVA and be at risk to die when he has high values of FBG and LPa laboratory tests together according to rule number 1. The patient will be at risk when he has a high value of HDL according to rule number 2. The results also show that the high value of CRP, TG will be affected according to rules number 3, 4, 5.

The learning accuracy of the model was 91%, since the model success to train about 273 example correctly from 300 examples as show at Figure 6.14. In addition to that we compared the rules inducted from ontological data mining algorithms with rules generated from other model, to make sure of success training of the algorithm and the results were positive.

6.6 Recommendations

The evaluation of each data mining methods represented in Table 6.4. We can see that association rule model has confidence value about 67.7 according to this if you are looking for good confidence use association rule. If you are looking at f-measure value and high learning accuracy we recommended decision tree model. Ontological data mining algorithm was the best learning accuracy.

Table 6-4 Alogorithms Evaluations Results

Model	Success learning accuracy	Confidence	f-measure
Association Rules	Na	67.7	Na
Decision Tree	88.25	Na	88.77
Rule Induction Data Mining	80.75	Na	83.84
Ontological Rule Induction	91%	Na	Na

Using ontology in the first data mining stages, data understanding and data prepressing and continued until the end of exploration helped in understanding the

results and reducing the number of rules and leafs in the association rules and decision tree model. Ontological data mining algorithm results achieved the goal of learning with high accuracy. The ontological data mining algorithm needs more work in order to conduct different evaluation methods, which is not suitable to the existing version of this algorithm. The evaluation of the generated results of all methods was very satisfying as shown in Table 6.4. The most important of the medical analysis is LPa, CRP, TG, FBG, Chol, HDL and the rise in the result of these analyses to diabetes patients may lead to cardiovascular disease. These results will be added to the proposed ontology to be determinant for cardiovascular disease occurrence among diabetes patients.

The benefit of the results obtained from this study will reduce the risk of developing cardiovascular disease and blood vessels in patients with diabetes, through the periodic follow-up of diabetes patients and the necessary tests have been concluded. Conducting these tests periodically determine the risk and thus driving diabetics to change their lifestyle to reduce danger and prevent disease.

The results of this study are important and novel, so we recommend that the primary health centres and clinics to be taken during the periodic monitoring of diabetics to reduce the possibility of cardiovascular disease and thus reduce the health cost of the Ministry of Health.

6.7 Summary

The preceding chapter have presented the results and explanations of different mining algorithms. Different evaluation methods are conducted to evaluate the models like accuracy, f-measure, confidence and learning accuracy. The recommendation resulted from the experiments are illustrated and explained. In the next chapter we have to conclude our work and explain what we have to do in the future.

Chapter 7

Conclusion and Future Works

Chapter 7

Conclusion and Future Work

7.1 Conclusion:

The factors affect the incidence of cardiovascular disease among diabetics are the most important things that you need to study and follow up. According to these factors that cause the disease, practitioners can make recommendations.

This thesis examines the factors affecting diabetic patients that may cause cardiovascular disease and blood vessels leading to death. Various techniques such as ontology, data mining and ontological data mining processes are used to generate a complete knowledge base system to diagnose the disease.

Different ranges of medical analyses were used as key factors in the prediction and diagnosis processes. Some other factors like diagnosing and symptoms have been used in building Ontology .

The ontology techniques were used with data mining to take advantage of the ontology in the exploration. This approach reduce the semantic gaps during the stages of the data mining process. Using ontology facilitating the process and clarifying the outputs of all stages of knowledge exploration and clarify the outputs of all stages of KDD .

Several techniques of the data mining method used in this study such as decision tree, association rules, and rule induction method to predict the possibility of cardiovascular disease among diabetics. The ontology was used in the first data mining stages, data understanding and data prepressing and continued until the end of exploration to understand the results and analyse them through the knowledge gained from it. Ontology has a main role in reducing the number of rules derived from all methods of exploration.

We used ontological data mining algorithm, which aims to take advantage of ontology reasoner and inferences. In the exploration process, where knowledge base was conducted rather than empirical data. This method aimed to improve the process of knowledge mining by taking advantage of ontology.

The study found that there is a group of medical tests for diabetics may have an impact on the cause of CVD. The most important of the medical analysis is the analysis of LPa, CRP, TG, FBG, Chol, HDL and that the rise in the result of these analyses may lead to cardiovascular disease and blood vessels.

The results of all the trial tests were evaluated in different ways to verify their validity. Confidence was calculated in addition to accuracy and f-measure. The high ratio of accuracy was 88.2%. In addition, f-measure were 88.77%.

The results of the ontological data-mining algorithm were evaluated by comparing them with the data mining to be sure of the safety of the bases derived. The results were satisfactory in addition to the accuracy calculation, which reached 91%...

Simple tests on ontology building must be carried out to evaluate them. Where the reliance on the interview with the expert to evaluate the results were positive in addition to using a Reasoner to make sure they are correct.

The methodology established in this study is a novel in the process of knowledge extraction and prediction in medical fields. The methodology used the integration of ontology with data mining in the process of knowledge discovery. Then the results have been integrated in the ontology to improve the knowledge base. These results will be added as a determinant of Cardiovascular and intracranial diseases with the aim of building an integrated knowledge base. The integrated cardiovascular diagnostics knowledge base can be used to build expert systems and other things.

7.2 Future work:

In the future work, we are looking forward to enhance the developed ontology with new terms and concepts and convert it to heavyweight one, in order to improve the diagnosing process and ontological data mining results. Expert system will be

developed for assisting the doctors in diagnosing issues. This methodology can be implemented to other medical disease, to help in diagnosing these diseases at early stages. This methodology can be extended to framework to help the practitioners who are working in medical field in diagnosing.

References

Abburu, S. (2012). A survey on ontology reasoners and comparison. *International Journal of Computer Applications, 57*(17).

Agarwal, S. (2018). How Diabetes Affects Your Body (And What You Can Do About It). *diabetes*.

Al-Hamadani, B. T., & Alwan, R. F. (2015). An Ontology-Based Expert System for General Practitioners to Diagnose Cardiovascular Diseases. *Advances in Computational Sciences and Technology, 8*(1), 53-65.

Arsene, O., Dumitrache, I., & Mihu, I. (2011). Medicine expert system dynamic Bayesian Network and ontology based. *Expert Systems with Applications, 38*(12), 15253-15261.

Association, A. D. (2017). 2. Classification and diagnosis of diabetes. *Diabetes care, 40*(Supplement 1), S11-S24.

Atiya Kazi, P. D. T. K. (2014). An Ontology Based Approach to Data Mining. *International Journal of Engineering Development and Research (IJEDR), 2*(4), 4.

Badran, M., & Laher, I. (2012). Type II diabetes mellitus in Arabic-speaking countries. *International journal of endocrinology, 2012*.

Banaś, P. (2012). Using the Protégé environment for building ontology for automated communication system at sea. *Zeszyty Naukowe/Akademia Morska w Szczecinie*, 12-17.

Barton, A., Rosier, A., Burgun, A., & Ethier, J.-F. (2014). *The Cardiovascular Disease Ontology*. Paper presented at the FOIS.

Brusa, G., Caliusco, M. L., & Chiotti, O. (2006). *A process for building a domain ontology: an experience in developing a government budgetary ontology*. Paper presented at the Proceedings of the second Australasian workshop on Advances in ontologies-Volume 72.

Casey, G. (2018). Type 2 diabetes mellitus. *Kai Tiaki: Nursing New Zealand, 24*(4), 20-24.

Cešpivová, H., Rauch, J., Svatek, V., Kejkula, M., & Tomeckova, M. (2004). *Roles of medical ontology in association mining crisp-dm cycle*. Paper presented at

the ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa.

Collaboration, E. R. F. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet, 375*(9733), 2215-2222.

Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications, 47*(10), 44-48.

Domingos, P. (2007). Toward knowledge-rich data mining. *Data Mining and Knowledge Discovery*, *15*(1), 21-28.

Dou, D., Wang, H., & Liu, H. (2015). *Semantic data mining: A survey of ontology-based approaches.* Paper presented at the Semantic Computing (ICSC), 2015 IEEE International Conference on.

Durairaj, M., & Ranjani, V. (2013). Data mining applications in healthcare sector: a study. *International journal of scientific & technology research*, *2*(10), 29-35.

El-Sappagh, S., & Ali, F. (2016). *DDO: a diabetes mellitus diagnosis ontology.* Paper presented at the Applied Informatics.

Fernández-López, M., Gómez-Pérez, A., & Juristo, N. (1997). Methontology: from ontological art towards ontological engineering.

Gedzelman, S., Simonet, M., Bernhard, D., Diallo, G., & Palmer, P. (2005). *Building an ontology of cardio-vascular diseases for concept-based information retrieval.* Paper presented at the Computers in Cardiology, 2005.

Giaretta, P., & Guarino, N. (1995). Ontologies and knowledge bases towards a terminological clarification. *Towards very large knowledge bases: knowledge building & knowledge sharing*, *25*(32), 307-317.

Giunchiglia, F., & Zaihrayeu, I. (2009). Lightweight ontologies *Encyclopedia of Database Systems* (pp. 1613-1619): Springer.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2006). *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*: Springer Science & Business Media.

Gorskis, H., & Chizhov, Y. (2012). Ontology building using data mining techniques. *Information technology and management science*, *15*(1), 183-188.

Gottgtroy, P., Kasabov, N., & MacDonell, S. (2004). An ontology driven approach for knowledge discovery in biomedicine.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

id Items, T. Association Analysis: Basic Concepts and Algorithms.

Ilayaraja, M., & Meyyappan, T. (2015). Efficient data mining method to predict the risk of heart diseases through frequent itemsets. *Procedia Computer Science*, *70*, 586-592.

Jain, V., & Singh, M. (2013). Ontology based information retrieval in semantic web: A survey. *International Journal of Information Technology and Computer Science (IJITCS)*, *5*(10), 62.

Jović, A., Gamberger, D., & Krstačić, G. (2011). Heart failure ontology. *Bio-algorithms and med-systems*, *7*(2), 101-110.

Kaushik, S., Choudhury, A., Dasgupta, N., Natarajan, S., Pickett, L. A., & Dutt, V. (2018). *Evaluating Frequent-Set Mining Approaches in Machine-Learning Problems with Several Attributes: A Case Study in Healthcare.* Paper presented at the International Conference on Machine Learning and Data Mining in Pattern Recognition.

Kolias, V. D., Stoitsis, J., Golemati, S., & Nikita, K. S. (2014). Utilizing semantic web technologies in healthcare *Concepts and Trends in Healthcare Information Systems* (pp. 9-19): Springer.

Kreinovich, V. (2017). *Uncertainty Modeling: Dedicated to Professor Boris Kovalerchuk on his Anniversary* (Vol. 683): Springer.

Kuo, Y.-T., Lonie, A., Sonenberg, L., & Paizis, K. (2007). *Domain ontology driven data mining: a medical case study.* Paper presented at the Proceedings of the 2007 international workshop on Domain driven data mining.

Larsson, S. C., Wallin, A., Håkansson, N., Stackelberg, O., Bäck, M., & Wolk, A. (2018). Type 1 and type 2 diabetes mellitus and incidence of seven cardiovascular diseases. *International journal of cardiology*, *262*, 66-70.

Lavrač, N., & Novak, P. K. (2013). Relational and Semantic Data Mining for Biomedical Research. *Informatica*, *37*(1).

Ławrynowicz, A., & Potoniec, J. (2014). Pattern based feature construction in semantic data mining. *International Journal on Semantic Web and Information Systems (IJSWIS), 10*(1), 27-65.

Liu, H., Dou, D., Jin, R., Lependu, P., & Shah, N. (2013). *Mining biomedical ontologies and data using RDF hypergraphs.* Paper presented at the Machine Learning and Applications (ICMLA), 2013 12th International Conference on.

Lovrencic, S., & Cubrilo, M. (2008). *Ontology evaluation-comprising verification and validation.* Paper presented at the Central European Conference on Information and Intelligent Systems.

Lucélia P. Branquinho, M. B. A., Renata M.A. Baracho. (2015). *Ontologies in support of data mining based on associated rules: a case study in a diagnostic medicine company*. Paper presented at the Brazilian Ontology Research Seminar, São Paulo, Brazil.

Manogaran, G., Varatharajan, R., & Priyan, M. (2018). Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. *Multimedia tools and applications, 77*(4), 4379-4399.

Masethe, H. D., & Masethe, M. A. (2014). *Prediction of heart disease using classification algorithms.* Paper presented at the Proceedings of the world Congress on Engineering and computer Science.

Medicine, J. H. (2018). Glossary - Cardiovascular Diseases. Retrieved from https://www.hopkinsmedicine.org/healthlibrary/conditions/cardiovascular_diseases/glossary_-_cardiovascular_diseases_85,P00222

MOH. (2015). *Health Status in Palestine*. Retrieved from Ministry of Health: Annual Report

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., . . . Fullerton, H. J. (2016). Executive summary: heart disease and stroke statistics-2016 update: a report from the American Heart Association. *Circulation, 133*(4), 447-454.

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology: Stanford knowledge systems laboratory technical

report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, Stanford, CA.

Ogurtsova, K., da Rocha Fernandes, J., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N., . . . Makaroff, L. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes research and clinical practice, 128*, 40-50.

Packer, M. (2018). Heart Failure: The Most Important, Preventable, and Treatable Cardiovascular Complication of Type 2 Diabetes. *Diabetes care, 41*(1), 11-13.

Panov, P., Džeroski, S., & Soldatova, L. (2008). *OntoDM: An ontology of data mining.* Paper presented at the Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on.

Parthiban, G., & Srivatsa, S. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS), 3*, 2249-0868.

Rajak, A., & Gupta, M. K. (2008). *Association rule mining: applications in various areas.* Paper presented at the Proceedings of International Conference on Data Management, Ghaziabad, India.

Ray, R. (2018). Advances in Data Mining: Healthcare Applications.

Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications* (Vol. 69): World scientific.

Seema, M. R., Mamta. (2012). Decision Tree: Data Mining Techniques *International Journal of Latest Trends in Engineering and Technology (JULIET), 1 Issue 3 September 2012*, 5.

Walters, J. (2008). Cardiovascular Disease (CVD).

Welty, C. (2003). Ontology research. *AI magazine, 24*(3), 11.

Whelton, P. K., Appel, L. J., Sacco, R. L., Anderson, C. A., Antman, E. M., Campbell, N., . . . Jessup, M. (2012). Sodium, blood pressure, and cardiovascular disease: further evidence supporting the American Heart Association sodium reduction recommendations. *Circulation*, CIR. 0b013e318279acbf.

Wu, C.-A., Lin, W.-Y., Jiang, C.-L., & Wu, C.-C. (2011). Toward intelligent data warehouse mining: An ontology-integrated approach for multi-dimensional association mining. *Expert Systems with Applications, 38*(9), 11011-11023.

Youn, S., & McLeod, D. (2006). Ontology development tools for ontology-based knowledge management *Encyclopedia of E-Commerce, E-Government, and Mobile Commerce* (pp. 858-864): IGI Global.

Appendix A

Domain Expert Ontology Evaluation Questions:

Three experts in the CVD domain works in Gaza Strip hospitals evaluated the proposed CVD ontology. The grades in the Table A-1 range from 1 to 5, where 1=strongly disagree, 2=disagree, 3=no preference, 4=agree, 5=strongly agree.

Table A.1 Domain Expert Ontology Evaluation Questions

Statements to evaluate the usability of the CVD ontology		DE1	DE2	DE3
1	I find the ontology easy to understand	5	3	4
2	I found that various concepts in this ontology were well integrated	5	5	5
3	I thought there was too much inconsistency in this ontology	1	1	1
4	I would imagine that most domain experts would understand this ontology very quickly	4	5	4
5	I am confident I understand the conceptualization of the ontology	5	4	4
6	I found a suitable number of classes and needed for the CVD domain ontology	4	4	4
7	I needed to ask many questions before I could understand the conceptualization of the ontology.	2	3	2
8	Class Laboratory tests is connected in proper way	5	5	5
9	The information doesn't enable proper conclusion	1	1	2
10	I think that I could contribute to this ontology	5	5	5
	Grades	Excellent	Excellent	Excellent

Appendix B

Domain Expert Results Evaluation Questions

Three experts in the CVD domain works in Gaza Strip hospitals evaluated the results of data mining algorithms. The Domain experts have proved the correctness of the generated rules as shown in the Table B-1. The grades in the table range from 1 to 5, where 1=strongly disagree, 2=disagree, 3=no preference, 4=agree, 5=strongly agree.

Table B.1 Domain Expert Results Evaluation Questions

Statements to evaluate the results of data mining algorithms		DE1	DE2	DE3
1	I find the attributes of the data set in enough	5	4	4
2	Dataset is clear and correct	5	5	5
3	There is not needed attributes	1	1	1
4	The rules generated are correct	5	5	5
5	the selected laboratory tests are enough for diagnosing CVD	5	5	5
6	LPa is the most important factor for CVD happening among diabetes	5	4	4
7	TG, CRP and FBG are affected to conducted CVD	5	5	5
8	There are other affected factors	4	3	4
9	We can recommend these results to be taken into consideration in MOH	5	5	4
10	I think that I could contribute to this work	5	5	5
	Grades	Excellent	Excellent	Excellent