Using Data Mining and Mobile Computing for Patients of Diabetes Disease: Case Study on Gaza Strip

استخدام تنقيب البيانات والحوسبة المتنقلة لمرضى السكر دراسة حالة قطاع غزة

By
Khalid S. Alsayed

Supervised by
Prof. Alaa El-Halees
Prof. of Computer Science

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Information Technology

April 2018
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.

<table>
<thead>
<tr>
<th>Student’s name:</th>
<th>Khalid S. Alsayed</th>
<th>اسم الطالب:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signature:</td>
<td></td>
<td>التوقيع:</td>
</tr>
<tr>
<td>Date:</td>
<td></td>
<td>التاريخ:</td>
</tr>
</tbody>
</table>
نتيجة الحكم

بناءً على موافقة عمادة البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على طرودة الباحث خالد سليم حنّسي السيد للدكتوراه في كلية منتجات المعلوماتقسم تكنولوجيا المعلومات وموضوعها:

استخدام تكنولوجيا البيانات والحاسوب المتصلة لمرضى السكر دراسة حالة عاطل غزة

Using Data Mining and Mobile Computing for Patients of Diabetes Disease:
Case Study on Gaza Strip

وعدد المناقشة والتي تمت اليوم الأحد 4 رمضان 1439هـ الموافق 2018/05/20م الساعة الحادية عشرة صباحا في قاعة الاجتماعات الكلية لجامعة الحكم العلمي والدكتوراه:

أ.د. علاء مصطفى الهلبو

إ.د. حسن الشامي

د. محمد عبد العظيف راشد

وبعد المناقشة أوصت لجنة بمنح الباحث درجة الماجستير في كلية تكنولوجيا المعلومات قسم تكنولوجيا المعلومات.

واللجنة لا تمنح هذه الدرجة إلا إذا توجب على الباحث تلبية الشروط القانونية والمطلوبة.

ووفقاً

عمادة البحث العلمي والدراسات العليا

أل. د. مازن عثمان
الموضوع/ استلام النسخة الإلكترونية لرسالة علمية

قامت إدارة المكتبات بالجامعة الإسلامية باستلام النسخة الإلكترونية من رسالة الطالب جلال سلحان حمزة.

رقم جامعي: 454212842338-

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

- تم إجراء جميع التعديلات التي طلبتها لجنة المناقشة.
- تم توقيع المشرف والمشرفين على النسخة الورقية لاعتمادها كنسخة معدلة ونهائية.
- تم وضع ختم "عمادة الدراسات العليا" على النسخة الورقية لاعتماد توقيع المشرف والمشرفين.

(Word PDF) وآخر (PDF)

(Word PDF +PDF) موجود جميع فصول الرسالة مجمعة في ملف وآخر

وجود فهرس الرسالة، والملخصين باللغتين العربية والإنجليزية بملفات منفصلة.

- تطبيق النص في كل صفحة ورقية مع النص في كل صفحة تقابلها في الصفحات الإلكترونية.
- تطبيق التنسيق في جميع الصفحات (نوع وحجم الخط) بين النسخة الورقية والإلكترونية.

ملفة: ستقوم إدارة المكتبات بنشر هذه الرسالة كاملاً بصيغة (PDF)

واحدة والترقية.

توقيع الطالب

19/9/2018
Abstract

International Diabetes Federation (IDF) in 2015 indicated that there are 415 million people living with diabetes, Type 2 Diabetes (T2D) is more common condition of diabetes types. The number is expected to increase to 642 million by 2040.

The lack of regular exercise, rising obesity rates, more munching food and smoking cigarettes are contributing factors to diabetes disease. Leading to an increased risk of infection of developing kidney disease, blindness, nerve damage, blood vessel damage and it contributes to heart disease.

Chronic care of diabetes comes with large amounts of data concerning the self and clinical management of the disease, so it is important to use the methods of data mining for analyzing and extracting knowledge from datasets of T2D patients. Also, with the increased popularity of smart phones, an increase in high-tech solutions has occurred especially in health care systems.

Many researchers proposed different solutions on how to use a mobile application by diabetes patients, however these researches neglects the status of each patient individually, other researchers used data mining methods to extract knowledge from patients data without the benefit of them by patients.

In this research, we propose a data mining system integrated with a mobile framework to extract new knowledge from diabetes healthcare data. The system aims to show the complications that a diabetic patient is likely to get based on the data he/she enters into the mobile system and the rules extracted from data mining system. The diabetes patients’ data were collected from UNRWA Clinics in Gaza Strip.

Results for the proposed data mining system showed that the average accuracy of classify new patients is 97.07%, in terms of the probability of infection of complications of the disease. In addition, the results of the questionnaire analysis about the idea of using the mobile application and the desire to use similar systems to diagnose other chronic diseases showed that the average specialists and patients' acceptance rate is equal to 81.64%.

Keywords: Data Mining, Diabetes Disease, Type 2 Diabetes, Mobile
ملخص الدراسة

أشار الاتحاد الفيدرالي الدولي لمرض السكري (IDF) في تقريره للعام 2015 أن هناك (415) مليون شخص في العالم مصابون بمرض السكري النوع الثاني (T2D)، وهو الحالة الأكثر شيوعاً بين أنماط مرض السكري المعروفة، ويتوقع لهذا الرقم أن يزداد ليصل إلى (642) مليون شخص بحلول العام 2040.

وفي هذا السياق، تعتبر قلة الممارسة المنتظمة للتمارين الرياضية وزيادة معدلات السمنة والتهام الطعام بكميات كبيرة والتدخين من أهم الأسباب المؤدية للإصابة بمرض السكري، والتي تؤدي بدورها إلى زيادة خطر الإصابة بأمراض متنوعة تشمل أمراض الكلى وفقدان الرؤية وتلف الأعصاب وتضرر الأوعية الدموية، بالإضافة إلى أمراض القلب.

إن التوصل إلى علاج دائم لمرض السكري تستوجب البحث في كميات هائلة من البيانات المتعلقة بكيفية إدارة المرض بواسطة الفراد أنفسهم أو بالعناية الطبية. لهذا المهم استخدام أساليب التنقيب عن البيانات لتحليلاً وتحليلماً قريباً من مرضى السكري من النوع الثاني. ويفيد ذلك مدى الشعبيات الكبيرة التي تتمتع بها أجهزة الاتصال الذكية التي توفر حلولاً تقنية عالية المستوى خاصة في مجال أنظمة عناية بالصحة.

وفي هذا السياق، فقد قدم العديد من الباحثين حلولاً متنوعة لمرضى السكري حول كيفية الاستفادة من تطبيقات الهواتف الذكية، لكن تلك الحلول تبقى مقتصرة لأن الباحثين تجاوزوا الحالة الخاصة بكل مريض. بعض الباحثين استخدموا أساليب التنقيب عن البيانات للوصول إلى نتائج من بيانات المرضى بدون فائدة ملموسة تعود عليهم.

يقترح القائمون على هذا البحث انشاء نظام تنقيب عن البيانات يتوافق مع بيئة الهواتف الذكية للوصول إلى نتائج جديدة من بيانات العناية الصحية بمرضى السكري. يهدف هذا النظام إلى تبيان مضاعفات المرض التي من المرجح أن يواجهها مريض السكري اعتماداً على البيانات التي يقوم بإدخالها في نظام الهاتف، بالإضافة إلى القواعد المستخرجة من نظام التنقيب عن البيانات، علمياً أن بيانات مرضى السكري في هذا البحث جمعت من العيادات الطبية التابعة لوكالة الفوغ في قطاع غزة.

آشرت نتائج التقييم لنظام التنبؤ عن البيانات المقترح أن متوسط الدقة عند تصنيف المريض الجديد بشأن احتمالية اصابته بمضاعفات المرض بلغ (97.07%). كما بنت نتائج تحليل الاستبان حول فكرة استخدام تطبيقات الهواتف الذكية والرغبة في استخدام أنظمة مشابهة لتشخيص أمراض مزمنة أخرى أن درجة موافقة المرضى والمختصين بهذا الشأن بلغت (81.64%).
كلمات مفتاحية: التنقيب عن البيانات، مرض السكري، مرض السكري النوع الثاني، الهواتف النقالة.
Dedication

This thesis is dedicated

To my parents who have always loved me unconditionally

To my sisters and brothers

To my wife, Sohad

To my son, Qosai

To my Daughters, Raghad and Dareen
Acknowledgment

All praise is to Allah

Many thanks and sincere gratefulness goes to my supervisor Prof. Alaa El-Halees, without his help, guidance, and continuous follow-up; this research would never have been.

I would like to extend my thanks to the academic staff of the Faculty of Information Technology who helped me during my master’s study and taught me different courses.

I would like to thank the officials in UNRWA Clinics in Gaza Strip for full collaboration.

Finally, I am greatly indebted to my family for their support during my course studies and during my thesis work.
Table of Contents

Declaration......................................................................................................................... I
Dedication ........................................................................................................................ VI
Acknowledgment .............................................................................................................. VII
Table of Contents ............................................................................................................ VIII
List of Tables .................................................................................................................... X
List of Figures ................................................................................................................... XI
Chapter 1 Introduction ...................................................................................................... 1
  1.1. Diabetes Disease ................................................................................................. 2
  1.2. The problem statement ...................................................................................... 4
  1.3. Objectives ............................................................................................................ 4
  1.4. The importance of research ............................................................................... 5
  1.5. Scope and Limitation .......................................................................................... 5
  1.6. Research Methodology ......................................................................................... 5
  1.7. Thesis Structure: ................................................................................................. 7
Chapter 2 Theoretical Foundation ................................................................................... 9
  2.1. Data Mining ......................................................................................................... 10
  2.2. Classification: ..................................................................................................... 14
  2.3. Mobile Application .............................................................................................. 19
  2.4. Summary ............................................................................................................ 19
Chapter 3 Related works ............................................................................................... 20
Chapter 4 Research Methodology ................................................................................... 28
  4.1. Collect and preparing the data set ........................................................................ 30
  4.2. Data mining methods .......................................................................................... 33
  4.3. Evaluate the data mining system ........................................................................ 41
  4.4. Develop mobile application ................................................................................ 44
  4.5. Evaluate the mobile application .......................................................................... 49
  4.6. Summary ............................................................................................................ 50
Chapter 5 Experiments and Results ................................................................................ 51
  5.1. Experiments Setup: ............................................................................................. 52
5.2. Results of the data mining model ................................................................. 52
5.3. Evaluation results of the data mining model ............................................... 68
5.4. Evaluation results of the mobile application questionnaire: ..................... 70
5.5. Summary ...................................................................................................... 73

Chapter 6 Recommendation and Future Works .............................................. 74
6.1. Thesis Summary .......................................................................................... 75
6.2. Conclusions ................................................................................................ 75
6.3. Recommendations ...................................................................................... 76
6.4. Future Works ............................................................................................. 77

The Reference ................................................................................................... 78
List of Tables

Table 4.1: Diabetic patients’ instance counts ................................................................. 30
Table 4.2: Diabetic patient’s data set attributes ........................................................... 30
Table 4.3: Diabetic complication attributes ................................................................. 32
Table 4.4: The mobile application questionnaire questions ......................................... 49

Table 5.1: Myocardial Infarction (MI) rules description ............................................... 53
Table 5.2: Congestive Heart Failure (CHF) rules description ...................................... 56
Table 5.3: Stroke rules description .............................................................................. 58
Table 5.4: End Stage Renal Failure (ESRF) rules description ..................................... 61
Table 5.5: Blindness rules description ......................................................................... 64
Table 5.6: Amputation rules description ..................................................................... 66
Table 5.7: Accuracy results of the data mining system ................................................. 69
Table 5.8: Mobile application respondent’s answers ................................................... 71
Table 5.9: Mobile application respondent’s acceptance ratio ...................................... 71
List of Figures

Figure 1. 1: The steps of approach about using data mining............................................. 7

Figure 2. 1: The world is data rich but information poor (Jiawei Han, 2012) ...................... 10
Figure 2. 2: Data mining - searching for knowledge in data (Jiawei Han, 2012) ............. 11
Figure 2. 3: The Process of Knowledge Discovery in Databases (Oded Maimon, 2010)..... 12
Figure 2. 4: The Data Classification Process ........................................................................ 15
Figure 2. 5: A decision tree for the mammal classification problem (Tan, 2006) .......... 17

Figure 4. 1: The steps of approach about using data mining............................................. 29
Figure 4. 2: The main process of decision tree method in RapidMiner tool for Myocardial Infarction (MI) complication............................................................... 34
Figure 4. 3: Change “MI” attribute to a class label.............................................................. 36
Figure 4. 4: Split ratio ......................................................................................................... 36
Figure 4. 5: Decision Tree method..................................................................................... 37
Figure 4. 6: The main process of decision tree method in RapidMiner tool for Congestive Heart Failure (CHF) complication ............................................................................. 37
Figure 4. 7: Change “CHF” attribute to a class label.......................................................... 37
Figure 4. 8: The main process of decision tree method in RapidMiner tool for Stroke complication.................................................................................................................. 38
Figure 4. 9: Change “Stroke” attribute to a class label....................................................... 38
Figure 4. 10: The main process of decision tree method in RapidMiner tool for End Stage Renal Failure (ESRF) complication................................................................................ 39
Figure 4. 11: Change “ESRF” attribute to a class label...................................................... 39
Figure 4. 12: The main process of decision tree method in RapidMiner tool for Blindness complication............................................................................................................. 40
Figure 4. 13: Change “Blindess” attribute to a class label................................................... 40
Figure 4. 14: The main process of decision tree method in RapidMiner tool for Amputation complication................................................................................................. 41
Figure 4. 15: Change “Amputation” attribute to a class label............................................ 41
Figure 4. 16: Training set and test set (Bramer)................................................................... 42
Figure 4. 17: Confusion matrix ............................................................................................ 43
Figure 4. 18: Mobile application screenshot 1 .................................................................... 44
Figure 4. 19: Mobile application screenshot 2 .................................................................... 45
Figure 4. 20: Mobile application screenshot 3 ................................................................. 46
Figure 4. 21: Mobile application screenshot 4 ................................................................. 47
Figure 4. 22: Mobile application screenshot 5 .................................................................... 48

Figure 5. 1: Myocardial Infarction (MI) decision tree output result ..................................... 53
Figure 5.2: Congestive Heart Failure (CHF) decision tree output result ....................... 56
Figure 5.3: Stroke decision tree output result .................................................................. 58
Figure 5.4: End Stage Renal Failure (ESRF) decision tree output result ...................... 61
Figure 5.5: Blindness decision tree output result ............................................................. 64
Figure 5.6: Amputation decision tree output result ......................................................... 66
Chapter 1

Introduction
1. Introduction

Computerization has led society to generate large amounts of data in almost all fields like business, marketing, surveillance, science, economics, fraud detection, sports, medicine etc. There is often knowledge hidden in the data, which once extracted, it can be used for decision making for business expansion. In healthcare environment data can be a great asset to organizations, but they have to be first transformed into knowledge (Koh & Tan, 2011). The healthcare environment is perceived as being “information rich” but “knowledge poor” (Obenshain, 2004). There is a wealth of clinical and administrative data available within healthcare systems, however, there is a lack of effectual analysis tools to discover knowledge contained in the databases of these systems (DeGruy, 2000).

Knowledge Discovery in Database (KDD) refers to the “Non-trivial extraction of implicit previously unknown and potentially useful information about data” (Kaur & Wasan, 2006). Data mining is the core of KDD, which is using for explore large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods (Han et al., 2011; Joshi et al., 2010; Lee et al., 2000; Obenshain, 2004; Thuraisingham, 2000).

The importance of using data mining techniques has emerged because of the need of powerful tools to analyze a large amount of data and transform these data into useful information and knowledge which helps policy makers to take the right decisions.

The discovered knowledge in healthcare databases can be used by healthcare administrators to improve operations and quality of service. It can be also used by healthcare professionals to improve their medical practice and patient care (DeGruy, 2000; Kaur & Wasan, 2006; Obenshain, 2004; Wasan et al., 2006) and diagnosis of several diseases such as Diabetes (Porter & Green, 2009), Stroke (Panzarasa et al., 2009), Cancer (Li et al., 2004), and Heart Disease (Das et al., 2009).

1.1. Diabetes Disease

The diabetes disease is characterized by persistent hyperglycemia (i.e. increased blood glucose levels), Blood glucose is the main source of energy which comes from the food we eat. A hormone made by the pancreas called Insulin helps glucose from food get into our cells to be used for energy, sometimes our body doesn’t make any or enough insulin or doesn’t use insulin well. Glucose then stays in our blood and doesn’t reach cells. Over time, having a large amount of sugar in the blood causes serious complications to human health. Although diabetes is not treated, taking some necessary measures prevents these complications (NIDDK, 2017).

There are two main types of diabetes: Type-1 and Type-2 Diabetes.

- **Type-1 Diabetes (T1D):**

  T1D occurs when the body begins to attack and destroy beta cells in the pancreas which produce insulin, leading to a lower level of insulin production and secretion in the blood. Symptoms include: extreme thirst and hunger, frequent urination, fatigue, poor vision and general fatigue. Treatment of this type is given daily doses of insulin for life (Association; Diseases).
• **Type-2 Diabetes (T2D):**

  T2D is a disease linked to the body resistance to insulin hormone produced by the pancreas, a condition that occurs when fat, muscle, and liver cells do not use insulin to carry glucose into the body’s cells to use for energy. As a result, the pancreas produces more and more insulin to maintain blood glucose levels. Over time, the pancreas doesn’t make enough insulin when blood sugar levels increase, such as after meals. This type of diabetes occurs most often in middle-aged and older people (Association; Diseases).

**Diabetes Key Facts**

- The World Health Organization (WHO) estimates that globally, high blood glucose is the third highest risk factor for early deaths, after high blood pressure and tobacco use (Organization, 2009).
- As of 2015, an approximately 415 million people had diabetes all over the world. This number is expected to increase to 642 million in 2040 (IDF, 2015b).
- T2D is a more common condition of diabetes types. In most countries, T2D has increased because of a rapid cultural and social changes, aging populations, increasing urbanization, physical inactivity, etc. (UN, 2002).
- In high-income countries, about 87% to 91% of all people with diabetes are estimated to have T2D, 7% to 12% are estimated to have T1D and 1% to 3% to have other types of diabetes (Evans et al., 2000).
- Many people remain undiagnosed because there are often few symptoms during the early years of T2D or symptoms that do occur may not be recognized as being related to diabetes (Dall et al., 2014).
- During the period from 2012 to 2015, approximately 1.5 to 5.0 million deaths each year resulted from diabetes (IDF, 2015c).
- The global economic cost of diabetes in 2014 was estimated to be 612 billion US Dollar, most countries spend between 5% and 20% of their total health expenditure on diabetes (IDF, 2013, 2015a).

**Diabetes’ challenges**

  Contributing factors to diabetes disease are a widespread lack of regular exercise, rising obesity rates, more munched food and smoking cigarettes. Diabetes increases the risks of developing kidney disease, blindness, nerve damage, blood vessel damage and it contributes to heart disease.

  Self-monitoring of blood glucose levels is an important factor in the management of insulin for diabetes patients. Through self-monitoring patients can help to prevent low blood sugar levels and high blood sugar levels occurring (Nguyen et al., 2012).

  A diabetic patient is suggested to take measurements of their blood glucose levels at least 3~4 times each day (Association). Preferably, these tests should be conducted before and after meals when possible. Paired measurements of blood glucose taken before and after meals serves to show if the patients’ blood glucose levels have been stable. In addition, it can also help medical professionals to identify
potential reasons of fluctuation. Measurement values are usually recorded in a logbook for reference when returning to a health institution for periodic checkups.

This leads to several challenges towards diabetic care. One is that patients may not always take measurements as required and properly record measured values. Another is that medical professionals only have access to the recorded measurements when the patient returns for periodic checkups.

With the increased popularity of smart phones, based on its mobility and wide range of usability that led it to become the fastest way for communicating, and increase in high-tech solutions has occurred, many smart phone users have access to applications which allow them to record their blood glucose results, meals and other factors which may then be used by the health care professionals to follow up their patients. Most of the applications lack an explanation about how they generate their solutions (Nguyen et al., 2012).

The proposed information system aims at supporting clinicians in the management of T2D patients, whom are the majority of diabetes patients and suffering from long term complications. The system provides personalized diabetes healthcare services for patients on smart phones by applying methods for data analysis, extraction of new medical knowledge using data mining algorithms, and developing a mobile application for show complication results.

1.2. The problem statement

Similar to many other chronic diseases, diabetes requires special care, patients require education on self-care such as blood-sugar monitoring, adherence to recommendations on diet, exercise, and daily lifestyle. Complications caused by badly managed diabetes create even more economic and emotional burdens on an already difficult condition, many patients have a tendency to disregard advice given by physicians or simply forget to take medication or measurements.

In addition, taking blood glucose measurements when visiting the clinic only may give incorrect results to specialists because the patient may improve his lifestyle or refrain from eating saturated fat and sugars before going to the health clinic.

Also, in health care centers there are data available which used only for daily transaction. It can be used in service improvement such as complications of diabetic patients.

Therefore, there is a need for mobile application to help diabetes patients in their daily life in order to avoid any complications.

1.3. Objectives

1.3.1. Main objective

The main objective of this thesis is to integrate data mining system that extract new knowledge from diabetes healthcare data, with a mobile application to show possible complications that a diabetic patient is likely to get based on the data he/she enters into the mobile application and the rules extracted from data mining system.
1.3.2. Specific objectives

- To collect diabetes patients’ data from health care institutes.
- To apply data mining methods (e.g. classification) to the dataset.
- To evaluate the classifier using testing data, to get the best accuracy.
- To use a rule extractor methods to extract the output rules from data mining system.
- To develop a mobile application that functions as an indicator of possible diabetes complication based on the mining results.
- To evaluate the usability of the system using users (patient) and experts (doctors).

1.4. The importance of research

People become diabetic patients because a lot of bad habits for long periods of time such as lack of exercise sports, eating unhealthy foods, and other reasons. Therefore, the goal of the development of intelligent systems based on the data mining algorithms and mobile application is to know the possible complications of chronic diseases, diabetes in our case, based on real data.

Our study also makes contributions to science, healthcare, and the society at large in addition to the direct support it renders at the individual level to prevent complications of diabetes, which, in addition to their negative health effects, affect the social and psychological situation of the patient and those around him.

1.5. Scope and Limitation

In this research we have some of the limitations such as:

- We selected dataset from UNRWA Clinics (Al Remal, Al Saftawi, Al Nusairat, Khan Younis, Al Shabora) in Gaza Strip.
- We used dataset already present for the year 2015.
- The target group of data those for patients with T2D because the highest percentage of patients belong to this type of disease, and this type cause greater complications on patients’ health.
- We will use existing algorithms in data mining like Classification algorithms such as Decision Trees.
- Develop an application running on a smart phones using Android operating system.
- The data mining server will be separated from mobile application for this stage.
- Security on the mobile application is out of our scope.

1.6. Research Methodology

Handling diabetes data passed through the following steps, as shown in Figure 1.1:

1. The dataset collected from UNRWA Gaza clinics as CSV files and imported into database, to link all the data related to a patient.

The dataset consists of patient’s personal information, medical history, disease type, Follow-up data, and treatment types and risk factors like age, smoking.
physical activity, family history and other factors, also the dataset consists of late complication that may infect the patient.

2. Prepared the data involves the following activities:
   - **Data Cleansing**: by handling the noise and missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.
   - **Relevance Analysis**: Dataset also has the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
   - **Data Transformation by using normalization method**: Normalization involves scaling all values for given attribute in order to make them fall within a small specified range.

3. Split dataset into training sample (2/3) and test sample (1/3).

4. Building the classifier using classification algorithms, the classifier is built from the training set made up of the dataset attributes and their associated class label(s).

5. System evaluation (metric) for data mining system: In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data if the accuracy is considered acceptable.

6. The output of the classifier is prediction of late complication type for patient.

7. We used rule extractor Java API routines to implement the results out from data mining system in the mobile application.

8. The mobile application developed by using Android Studio IDE, the application contained all factors that effect on diabetes complication. The interface is in Arabic language since some patients may not understand English.

1.7. **Thesis Structure:**

This thesis consists of six main chapters, which are structured around the objectives of the research. The main points discussed throughout the chapters are listed below:

- **Chapter 1: Introduction:** It gives a short introduction about the diabetes disease, the thesis problem and objectives.
- **Chapter 2: Theoretical Foundation:** This chapter presents an overview of data mining techniques and the steps of data mining life cycle process, explains data mining methods, focus on classification methods. Finally, we focus on mobile application.
- **Chapter 3: Related Works:** presents other works related to the thesis.
- **Chapter 4: Research Methodology:** This chapter explains the proposed approach about using data mining and mobile application for patients of diabetes disease. An explanation about the data sets used in the experiments,
preprocessing of these data set and finally the evaluation methods for the data mining system and the mobile application.

- **Chapter 5: Experiments and Results:** gives in detail about the sets of experiments, analyzes the experimental results. In addition, it gives a discussion for each experiment and evaluation results.

- **Chapter 6: Recommendation and Future Works:** discusses the final conclusions and presents possible future works, finally gives some useful recommendations about the thesis.
Chapter 2

Theoretical Foundation
The aim of this thesis is to use data mining methods to extract knowledge of data and connect it with a mobile application. Therefore, this chapter present an overview of data mining techniques and the steps of data mining life cycle process, explain data mining methods, we focus on classification methods because we use it in our research. Finally, we focus on mobile application.

2.1. Data Mining

2.1.1. Introduction

Computerization has led society to generate large amounts of data in almost all fields like business, marketing, surveillance, science, economics, fraud detection, sports, medicine etc., Terabytes of data pour into our computer networks and various data storage devices every day, this phenomenal growth in data volume is the result of the computerization of our community and the rapid development of data collection and storage tools (Jiawei Han, 2012).

The abundance of data and the rapid growth in the volume of data stored exceed our human ability to understand or infer knowledge (Figure 2.1) without the use of powerful tools to extract knowledge. As a result, data collected in large data repositories is rarely visited, decisions are often made not based on data stored in data repositories but based on the intuition of the decision maker because the decision maker does not have the tools to extract the knowledge embedded in the vast amounts of data. (Jiawei Han, 2012)

Figure 2.1: The world is data rich but information poor (Jiawei Han, 2012)
Data mining which is also known as Knowledge Discovery in Database (KDD) refers to the “Non trivial extraction of implicit previously unknown and potentially useful information about data” (Kaur & Wasan, 2006). Data mining is the core of KDD, which is using for explore large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods (Figure 2.2) (Han et al., 2011; Joshi et al., 2010; Lee et al., 2000; Obenshain, 2004; Thuraisingham, 2000).

KDD used by researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The goal of the KDD process is to extract knowledge from data in the context of large databases.

2.1.2. The KDD Process:

The KDD process is iterative and interactive, a typical KDD process is consists of nine steps as seen in figure 2.3 (Thakare). The process is iterative at each step, meaning that moving back to adjust previous steps may be required. The process starts with determining the KDD goals, and ends with the implementation of the discovered knowledge. Following is a brief description of the nine-step KDD process (Oded Maimon, 2010):
1. Understanding the application domain: which prepares the scene for understanding what should be done with the many decisions. The people who are in charge of a KDD project need to understand and define the goals of the end-user and the environment in which the knowledge discovery process will take place. As the KDD process proceeds, there may be even a revision and tuning of this step.

2. Selecting a data set on which discovery will be performed: The data that will be used for the knowledge discovery should be determined by finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process. This process is very important because the data mining learns and discovers from the available data. If some important attributes are missing, then the entire study may fail.

3. Preprocessing and cleansing stage: includes data cleaning, such as handling missing values and removal of noise or outliers. This stage may involve complex statistical methods, or using specific data mining algorithm. For example, in case of a certain attribute is not reliable enough or has too many missing data, then this attribute could become the goal of a data mining supervised algorithm for predicting the missing data.

4. Data transformation: the goal of this stage is to find useful features to represent the data depending on the goal of the task, and use dimensionality
reduction or transformation methods to reduce the effective number of
variables under consideration or to find invariant representations for the data.

5. Choosing the appropriate data mining task: This mostly depends on the KDD
goals, and also on the previous steps which may be classification, regression,
clustering … etc.

6. Choosing the Data Mining algorithm: This stage includes selecting method(s)
to be used for searching for patterns in the data, deciding which models and
parameters may be appropriate.

7. Employing the Data Mining algorithm: Results to be satisfied, we might need
to employ the algorithm several times by tuning the algorithm parameters.

8. Evaluation: At this stage the patterns are evaluated and interpreted with
respect to the goals defined in the first step. This step focuses on the
comprehensibility and usefulness of the induced model. In addition, in this
step the discovered knowledge is documented for further usage.

9. Using the discovered knowledge: The success of this step determines the
effectiveness of the entire KDD process, this comes through the integration
the discovered knowledge into another system for further actions.

2.1.3. Data Mining Functionalities:

Because of the variety of data and information repositories on which data
mining can be performed, data mining functionalities are used to specify the kinds of
patterns to be found in data mining tasks. Such tasks can be classified into two
categories, descriptive and predictive. Descriptive mining tasks characterize
properties of the data in a target data set. Predictive mining tasks perform induction
on the current data in order to make predictions.

Data mining functionalities, are described below:

- **Classification:**
  
  Classification is a form of data analysis that extracts models (called
classifiers) that describe important data classes, also known as supervised
classification. Classification methods are used to predict categorical class
labels. Data Classification is a two-step process, the first step is a learning
step where a classifier is constructed, and the second step is a classification
step where the classifier is used to predict class labels for new data (Jiawei
Han, 2012).

- **Prediction:**
  
  Similar to classification, except that the constructed classifier predicts a
continuous-valued function, as opposed to a class label. Prediction is most
often referred to the forecast of missing numerical values, or increase/
decrease trends in time-related data. The major idea is to use a large number of past values to consider probable future values (Asghar & Iqbal, 2009; Jiawei Han, 2012)

- **Clustering:**

  Clustering is the process of partitioning a large data sets into groups according to their similarity. Each group is a cluster, such that each object in a cluster are similar to one another, and dissimilar to objects in other clusters. Clustering is also called unsupervised classification because the classification is not dictated by giving class a label(s) (Jiawei Han, 2012).

- **Association rules:**

  Association rule analysis is used to extract correlations, frequent patterns or associations within sets of items in the transaction databases or other data repositories. Association rule mining is to discover association rules that realize the predefined minimum support and confidence from a given database, whereas support refers to the number of transactions that contain a specific itemset, and confidence refers to the conditional probability that an item appears in a transaction when another item appears (Kotsiantis & Kanellopoulos, 2006).

- **Outlier detection:**

  Outlier detection is the process of discovering data objects with practices that are altogether different from desire, such objects are called outliers. In another meaning, an outlier is a data object that significantly different from the rest of the objects, as if it were generated by a different mechanism. Outliers are different from noisy data, noise is a random error or difference in a measured variable. In data mining tasks, noise should be removed before outlier detection (Jiawei Han, 2012).

**2.2. Classification:**

Classification is a form of data analysis that extracts models (called classifiers) that describe important data classes, also known as supervised classification. Classification methods are used to predict categorical class labels (Jiawei Han, 2012). For example, the classification model can be built to predict whether the risk of any insurance transaction is “high” or “low”. The classification methods can be implemented in a variety of domains such as marketing, manufacturing, and medical diagnosis.
Data Classification is a two-step process, the first step is a learning step (Figure 2.4-a) where training data are analyzed by a classification algorithm. In example of figure 2.4, the class label attribute is “Risk” that represent whether the insurance transaction is either “High” or “Low”. The classifier is represented in the form of classification rules. This step can also be represented as the function $Y=f(X)$, that can predict the value of class label “$Y$” of a given data “$X$” by applying the classification rules.

Since the class label of each training data is given, this step is also known as supervised learning (i.e., the learning of the classifier is “supervised” in that it is told to which class each training data belongs), which is different from unsupervised learning (or clustering) where the class label of each training data is not known, and the number or set of classes to be learned may not be known.

The second step is a classification step (Figure 2.4-b). The predictive accuracy of the classifier is estimated. The accuracy of a classifier is calculated by applied the classification rules that generated from training data on a given test data (The test data are independent of the training data, meaning that they were not used to build the classifier), the associated class label of each test data is compared with
the learned classifier’s class prediction for that data. The higher percentage of accuracy means that the higher correctly test data classified by the classifier (Jiawei Han, 2012; Joshi et al., 2010; Oded Maimon, 2010).

2.2.1. Classification Algorithms:

Many algorithms for data mining classification were presented over time such as Rule Based classifier, k-Nearest Neighbors and Decision Trees (Mitchell, 1997), in this section we will discuss some of these algorithms, and at the end of this section we will discuss in more details the decision trees algorithm because we used it in our research.

- Rule Based classifier:

This algorithm is one of Rule-based machine learning approaches (Urbanowicz & Moore, 2009). Rule based classifier uses a set of IF-THEN rules for classification, (e.g. IF 'condition' THEN 'conclusion'). The condition part consists of one or more attribute tests and these tests are logically ANDed, while the conclusion part consists of class prediction. When the condition part holds true for any given test data, then the conclusion part is satisfied. Evaluation of any rule can be done by calculating the coverage and the accuracy measures. The rule coverage is the percentage of data that satisfy conditions, where the accuracy of the rule represents the percentage of data that are correctly classified by a rule (the number of correct predictions divided by the rule coverage). Ideal rules should have both high coverage and accuracy rates (Jiawei Han, 2012; Qin et al., 2009).

- k-Nearest Neighbors:

K-Nearest neighbor is a supervised learning algorithm where the result of new instance query is classified based on the majority of K-nearest neighbor category. The purpose of this algorithm is to classify a new object based on attributes and training samples. K-Nearest neighbor is a type of lazy learning, where the function is just approximated locally and all calculation is deferred until classification (Phyu, 2009; Sutton, 2012).

To predict the classification of the new object:

1) First we determine a user defined parameter “k” which represent the number of nearest neighbors.
2) Calculate the distance between the new object and all the training samples.
3) Sort the distance and determine nearest neighbors based on the K-th minimum distance.
4) Gather the classification of the nearest neighbors
5) Use the simple majority of the classification of nearest neighbors as the prediction value of the new object.
• Decision Trees:

Decision tree is a method commonly used in data mining. Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing, production and financial analysis. The goal is to create a model that predicts the value of a class label based on several input variables. As shown in figure 2.5, a decision tree is a tree-structured plan where each internal node (non-leaf node) represents an attribute, each branch represents a value of the attribute and each leaf node (terminal node) holds a class label (Jiawei Han, 2012; Rokach & Maimon, 2014).

![Decision Tree Example](image)

**Figure 2.5:** A decision tree for the mammal classification problem (Tan, 2006)

To know how the decision tree used for classification, suppose a new object which its class label is unknown. The attribute values of the new object are tested against the decision tree. A path is traced from the root to a leaf node, which indicates the class label for that new object.

Most algorithms for decision tree induction follow a top-down approach, which starts with a training set of data and their related class labels. The training set is recursively divided into smaller subsets as the tree is being built. A basic decision tree algorithm is summarized in the following steps (Jiawei Han, 2012):

1) The algorithm is called with three parameters: “D”, *attribute list*, and *Attribute selection method*. Where “D” refers to a dataset (initially, it consists of a complete set of training data and their related class labels). The parameter *attribute list* is a list of attributes describing the data. The parameter *attribute selection method* specifies a heuristic procedure for
selecting the attribute that “best” distinguish the given data according to class, this procedure uses a criterion selection measure such as information gain.

2) Create a single node “N” which representing the training data in “D”.

3) If the data in “D” are all of the same class, then node “N” becomes a leaf and labeled with that class.

4) Otherwise, the algorithm calls criterion selection method to determine which attributes to be at node “N” by determining the “best” way to divide the data in “D” into individual classes, also tells us which branches to grow from node “N”.

5) The node “N” is labeled with the splitting criterion. A branch is grown from node “N” for each outcome of the splitting criterion. The data in “D” are divided accordingly.

6) The algorithm uses the same process recursively to form a decision tree for the data at each resulting partition, “Dj”, of “D”.

7) The recursive dividing stops when any of the following terminating conditions is true:
   • All the data in partition “D” belong to the same class.
   • There are no remaining attributes on which the data may be further partitioned. In this case, the node “N” converts into a leaf and labeling it with the most common class in “D”.
   • There are no data for a given branch, that is, a partition “Dj” is empty. In this case, a leaf is created with the majority class in D.

8) The resulting decision tree is returned.

To determine which attribute will be chosen to divide the data when creating a decision tree, an attribute selection measure provides a ranking for each attribute describing the given training data. The attribute having the best score for the measure is chosen as the splitting attribute for the given data. The tree node created for partition “D” is labeled with the splitting criterion, branches are grown for each outcome of the criterion, and the data are partitioned accordingly. Information gain is one of the popular attribute selection measures (Yang & Pedersen, 1997).

Information gain is an impurity-based criterion that uses the entropy measure as the impurity measure (Oded Maimon, 2010).

\[
\text{Information Gain}(a_i, S) = \sum_{v_i \in \text{dom}(a_i)} \frac{|\sigma_{a_i = v_i} \cdot S|}{|S|} \cdot \text{Entropy}(y, \sigma_{a_i = v_i} \cdot S)
\]

where:

\[
\text{Entropy}(y, S) = \sum_{c_j \in \text{dom}(y)} \frac{|\sigma_{y = c_j} \cdot S|}{|S|} \cdot \log_2 \frac{|\sigma_{y = c_j} \cdot S|}{|S|}
\]

(1)
2.3. Mobile Application

A mobile application is a type of application software designed to run on a mobile device such as smartphone or tablet computer. Mobile applications frequently serve to provide users with similar services to those accessed on personal computers.

Usage of mobile applications has become increasingly common by mobile phone users (Ludwig, 2015). Researchers found that usage of mobile applications correlates with user context and depends on user’s location and time of usage (Böhmer et al., 2011). Mobile applications are playing an increasing role within healthcare and can yield many benefits when designed and integrated correctly (Marcano-Belisario et al., 2016; Ventola, 2014).

Mobile applications come in two formats, native applications and mobile web applications. Here’s a brief description of each:

- **Native applications:**
  A native mobile application is a piece of software for smartphones and tablets. Native applications are built specifically for each mobile platform and installed on the device itself. Each native mobile application only works on the platform for which it was built.

- **Web applications:**
  A mobile web application is a web application formatted for use on a smartphone or tablet and accessed through the device’s web browser. Mobile web applications are platform independent since they are accessed through the browser without requiring installation on each device.

In our research, the mobile application was developed to work on Android devices. Android is an open source software package and Linux based operating system, initially developed by Android Inc. and then bought by Google in 2005 (Elgin, 2005).

The mobile applications are usually developing in Java programming language using the Android software development kit (SDK), Android Studio is the official Android integrated development environment (IDE).

2.4. Summary

In this chapter we presented the three main categories of theoretical foundation in our thesis, data mining, classification and mobile application. In data mining we presented the main aspects and functionalities in data mining, in classification we presented a general review of classification in data mining, brief description for some algorithms used classification techniques and finally, more details about decision tree algorithm that we used in our research. In the last section we presented the definition of mobile application and a brief description of its format.
Chapter 3
Related works
Similar to many other chronic diseases, diabetes requires multidisciplinary care, and patients require education on self-care such as blood-sugar monitoring, adherence to recommendations on diet and exercise. A growing body of evidence suggests that diabetes-management programs need an information technology backbone in order to be effective (Bu et al., 2007). From a health system perspective, high-quality data on disease trends, cost, and quality of care are vital to developing, monitoring, and evaluating diabetes prevention and control programs.

Increasing the computing power of high-end cell phones has a positive impact on increasing the access to the Internet. This low-cost communication platform is capable of addressing the data requirements of the health system and continued care for people with diabetes as well.

One of the earliest cases of provisioning healthcare with telecommunications technology was Jones, et al. (1998), who uses the Web to facilitate communications between healthcare providers and elderly diabetic. In this system, clinicians and patients were able to use a Web-browser based client to interact with each other and share data (Jones et al., 1998).

This chapter contains three parts:

**A) Diabetes with Data Mining and Mobile:**

Tsai, et al. (2012), bring together modern telecommunications technologies and healthcare expertise to provision a mobile framework for diabetes telehealth. They developed an Android application that functions as an electronic logbook to record the patient’s blood glucose measurements. A patient is diagnosed with diabetes must measure and record their blood glucose levels several times every day. Health education teaches the patient how to take a sample of his or her own blood and analyze the sample using a glucometer. The patient is then able to use the Android application on the smart phone to record the glucometer readings. Medical professionals are able to access and view data uploaded by patients they look after from a web-browser based client. The program, which was developed, depends only on the level of glucose in the blood, without considering the other factors that indicate the level of progress of the disease, such as sex, weight or age of the patient.

The primary objective of the program which was developed in the study is to collect information only as a first step for data mining operations without any use of the extracted rules in case of new record entered (Tsai et al., 2012).

Luo, et al. (2014), implemented a data-driven, data-mining based lifestyle management solution for T2D called GlucoGuide. GlucoGuide conveniently collects a variety of lifestyle data (diet, exercise, etc.) via medical sensors and wearable devices, and uploads the data to the computing server. If the data indicate that a
patient is in emergent or dangerous situations, such as abnormal high or low blood glucose levels, GlucoGuide provides immediate assistance to the patient, as well as alerts to healthcare team. If patients’ data do not present emergent or dangerous situations as above, data are collected and uploaded to the computing server. After that, a data mining framework deployed in the server to discover correlations between the recent lifestyle data and the blood glucose levels for each patient. Such correlations will be framed in natural language templates, and sent to patients’ smart phones as recommendations (Luo, Ling, Schuurman, et al., 2014).

Luo, et al. (2014), proposed, built, and evaluated a food classification tool using mobile computing and predictive models to proactively guide T2D patients along their diet selection, the predictive model classified each food item into three classes (e.g., “Choose More Often”, “In Moderate”, and “Choose Less Often”). The proposed tool contains three main components, a comprehensive and commercialized food database, a mobile application, and a predictive model for food classification. The database consists of large number of food items with its name, category, subcategory, and many nutrient attributes. The mobile application including Android UI components, voice searching, and Optical Character Recognition (OCR) module to facilitate the database search. The predictive model predicts whether or not any food item is suitable for patient diabetic condition (Luo, Ling, & Ao, 2014).

Priya, et al. (2015), aimed to build a model using data mining techniques and a mobile application to predict whether a person has diabetes or not. The researchers used the C4.5 classification algorithm on a Pima Indians Diabetes Database (PIDD) dataset. The mobile application allows the user to enter a set of related parameters, these data is compared with the rules generated from the use of the algorithm, the generated result indicate whether the person has diabetes or not (Priya et al., 2015).

Sittig and a team of researchers created a method for development a mobile application through consultations, health informatics, and data mining of diabetic data for 99 patients with type 2 diabetes to change their health outcomes. The developed application give recommendations for improving diabetes care, these tips include reminders to take the treatment, re-supply the medicine pills from the pharmacy and glucose monitoring. In addition, the mobile application could boost health outcomes by monitoring diet plan, adherence to exercise schedule and notify the patients with tips about change their lifestyle to improve wellness (mHealthIntelligence, 2015).

Machado, et al. (2017), introduced additions to an existing mobile application (MyDiabetes) by incorporate standard medical protocols, advice and directives to provide early guidance to avoid some of the complications associated with diabetes. In addition, they described an initial approach for using a data mining algorithms and methods to discover patterns that may cause critical situations.
The researchers aimed to provide diabetic patients and clinicians with a mobile assistant enable to collecting patients' data, then send feedbacks and advices to enhance diabetes management and changing the users' current treatments.

While the researchers relied on standard medical protocols to give advice and guidance to diabetics based on patient data to develop mobile application, they collected data of 31 volunteers to generate an interesting association rules in the data mining approach. These rules can be used to predict the probability of a user to have hypoglycemia or hyperglycemia according to certain conditions (Machado et al., 2017).

B) Diabetes and Mobile Applications:

Many researchers proposed different solutions on how to use a mobile application by diabetes patients. This subsection reflects a number of researches that worked on management diabetes through a mobile application.

Gittens et al. (2014), proposed a Mobile Health Consultation Application (MHCA) that provides daily tips on how to better manage diabetes and consultations on daily activities to improve regimen compliance of diabetics. The application allows the user to register and create a digital record for them. The account contains the user’s personal information and stores information about them that the application has collected and evaluated. The application also contains a “Mobile Consult” which consists of questions the user must answer about their daily activities based on the guidelines of the American Diabetes Association. The application also contains “The Tip of the Day” feature that provides the user with an important bit of information everyday which could be useful in managing their disease (Gittens et al., 2014).

Alanzi et al. (2014), presented a review of the status of Cognitive Behavioural Therapy (CBT) in diabetes management and proposed a model for the implementation of CBT into a mobile diabetes management system by using smart mobile phone. CBT is a psychotherapeutic approach that addresses different psychological problems, such as maladaptive behaviors, psychological disorders, stress, health conditions, etc.

Mobile Diabetes-Cognitive Behavioral Therapy (MD-CBT) model assists the patient to provide his/her CBT data remotely via the smartphone, it also assist the therapist by sending feedback and observing the state of the patient’s behavioral changes remotely.

The MD-CBT cycle starts from the event in which high blood glucose levels of the patients are recorded or observed from the diabetes management system, based on classification algorithm which is a software that triggers the MD-CBT therapy
module and the automated message, if the specified criterion is met. The system will automatically send an SMS to the patient, requesting the submission of his or her CBT data, the patient will then submit the CBT data, which comprises several choices about thought, mood, and action which are predefined in the system by the therapists. The submission data is then analyzed by the CBT therapist, who sends feedback to the patient accordingly (Alanzi et al., 2014).

Koutsouris et al. (2014), addressed the importance of using mobile telephones from both diabetics and professionals for diabetes monitoring and the techniques that will enable these systems to reduce the aggravation of disease complications. The researchers pointed out that the mobile phones could serve as a tool for collecting information that could be used for health monitoring and provide tips, which enables patients with diabetes to enhance their diabetes education.

The researchers also noted that the use of text messages could allow patients to log their health-related data and receive customized feedback based on that data. In addition, text messages can used to remind patients to log any health-related data, patients can reply to messages, and the system may process these messages to update the patients' personal record. Another use of mobile phones in diabetes care is remote monitoring, patients could send their daily data, such as blood glucose through mobile phones to professionals who provide advice back to patients by the same way.

Researchers also advice to build systems that combines a mobile phone application and a variety of Bluetooth-connected devices used for collecting and analyzing symptoms by sensing or self-report such as devices for measurement of blood pressure or blood glucose level which are automatically uploaded to a server and are continuously monitored. In case of any dangerous signs, then the patient’s physician immediately notified. This enables early detection and prevention of critical events (Koutsouris et al., 2014).

Al-Taee, et al. (2015), presented a mobile health approach that provides a multiple care dimensions of diabetes by means of remote collection and monitoring of patient data and provision of personalized and customized feedback on a smartphone platform and maintaining continuous interactivity between the patient and his/her health care team. Architecture of the platform comprises three layers; physical objects, network, and a remote web-based layer.

The physical objects layer performs the communications between people and device objects as well as between objects and other objects, it involves several medical sensors (blood glucose monitor, blood pressure, pulse rate monitor, and weight scale) and a smart phone device, all these devices communicates to each other via Bluetooth connectivity. The network layer represented by the long-range connectivity between the physical layer and the web-based layer using 3G and/or
Wi-Fi network linked to the Internet. The web-based layer represents the application layer of the platform. It interfaces the various objects of the physical layer to other objects (e.g., healthcare professionals, hospitals, and other systems). It is also responsible for remote data collection and storage, data processing and monitoring, and making decisions based on constraints specified by individual patient treatment plans.

The data collected by the platform is that information relevant to the daily self-management of the disease, including: diet, insulin intake, physical activity, and illnesses or Measurements from patient’s medical sensors via Bluetooth. The collected measurements and information are then uploaded to the web-based portal and presented in various ways such as tabular, graphical, and others.

The platform provides a Real-Time Decision Support like feedback the appropriate insulin bolus for patients of T1D. The proposed platform also facilitates connectivity between patients and their professional health cares to improve patients’ adherence to their individualized treatment plans (Al-Taee et al., 2015).

C) Diabetes and Data Mining:

Data mining is one of the ongoing domains in computer science, which studies applications of computational methodologies in real life. S Sankaranarayanan (Sriram Sankaranarayanan & Perumal, 2014) applied a Rule Set and Decision Trees classification methods on a dataset created by George John from a data repository at http://kddics.uci.edu to determine either a person is or not having diabetes mellitus. This study used data mining classification methods for diagnose new cases and do not make any interaction or communication between patients and medical specialists like our proposed method.

Brown, et al. (2013), their aim is to help explain suggested doses for diabetes patients, they used the Euclidean distance to determine the distance between numeric values. The returned cases from the knowledge base are those of closest similarity to the query used in retrieval, firstly a query is defined by asking the user a series of questions to build a picture of the new situation. Once a query representing the new problem has been defined, the similarity comparisons are run against the knowledge base. The application runs on desktop that does not give the flexibility to the patient to introduce new readings such as the use of mobile. Sensible boundaries need to be placed upon attribute values that may be input by the user and also on the provided data to prevent dangerous outcomes and maintain data integrity. The system should either ignore anomalies cases or notify the user of such cases for them to make an educated judgement. The user will be prompted to confirm the previous insulin recommendation provided an acceptable solution (Brown et al., 2013).
Sankaranarayanan et al. (2014), presented a survey of set of data mining methods for predicting immediate or later incidence of diabetes mellitus. They used two association rules algorithms, FP-Growth and Apriori on a diabetes dataset to generate useful rules used to improve the expert systems and to make better clinical decision making in India. The researchers applied the two association rules algorithms on a Pima Indians Diabetes Database (PIDD) dataset, which contains of 768 patients' data, each of them consists of nine attribute variables. Eight of these attributes are inputs and the last one being the output. The input attributes are number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index (BMI), diabetes pedigree function and age. The goal is to use these eight variables to predict the attribute value of the last one. The results of the research showed that both algorithms generates the same number of frequent sets and the same number of rules for same dataset under the same constraints (S Sankaranarayanan, 2014). While this research focus on predicting immediate or later incidence of diabetes mellitus by applying an association rules, our research focus on predicting the late complications of diabetics by implementing a data mining classification methods.

Vijayalakshmi et al. (2017), presented a study aimed at analyses the risk factors for diabetes by using data mining methods. They subjected the dataset to classification using two different decision tree induction methods and comparing the results of these methods, and then built a software to predict diabetics based on the knowledge extracted from the classification methods. They also generated a rules that govern diabetes using association rules mining. In addition, they subjected the dataset to clustering to perform descriptive data mining. The researchers applied the data mining methods on a dataset consists of 337 instances of patients' data. The data related to 202 diabetics and 135 non-diabetics.

To build the classification model, they used the Random Tree and J48 algorithms to classify the diabetic patient into two classes based on the value of HbA1c (accumulative blood glucose level for 3 months), under control (HbA1c < 7%) and out of control (HbA1c > 7%). The study showed that the J48 algorithm is better technique in terms of accuracy of classifying with an accuracy of 81%. In association rules mining, they selected the most 10 interesting rules from the set of all possible rules, by using minimum support value equal to 0.75 and 0.9 as minimum confidence. In clustering, they used the k-means algorithm. The result of clustering technique grouped 80 instances as non-diabetics and 257 as diabetics.

While this research focus on analysis of risk factors for diabetes by applying several data mining techniques, our research focus on predicting the late complications of diabetics by implementing a data mining classification methods (Vijayalakshmi & Jenifer, 2017).
Wu, Han, et al. (2017), presented a novel model for predicting type 2 diabetes mellitus based on data mining techniques, the study aimed to improve the accuracy of predicting and to produce a model adaptive to more than one dataset. The presented model consists of two parts, the K-means algorithm for clustering the data and the Logistic Regression algorithm for classification. In addition, they compared the results with the results of other researchers. First, the researchers applied the two algorithms on a Pima Indians Diabetes Database (PIDD) dataset, which contains 768 patients' data. The results of the comparison showed that the presented model gave a higher accuracy by 3.04% than other researchers for the same dataset with an accuracy of 95.42%. To demonstrate the prediction accuracy and adaptability of the model, the researchers applied the model to two other datasets. The results of the evaluation showed that the accuracy of the model prediction reached 90.7% and 93.5% (Machado et al., 2017).

While the researchers built a data mining model to predict whether the patient is type 2 diabetic or not by using a clustering and classification algorithms, we used a decision tree classification algorithm to predict the late complications of diabetes for persons who are already diabetics.

**Summary**

As presented in previous sections, some of the researches developed methods depends on provide some general tips or consults to diabetes patients on mobile application regardless of the status of each patient individually, or using data mining methods to extract knowledge from patients data without the benefit of them by patients. In contrast, our proposed method benefit from the results of using the data mining methods to monitor and give advice for each patient based on his/her health status.

Due to the difficulty of obtaining medical sensors and wearable devices because of the high price compared with the per capita income in the Gaza Strip, our proposed method will be only on mobile devices and this is the best solution because of the availability with the majority of patients.
Chapter 4

Research Methodology
This chapter explains our approach about using data mining and mobile application for patients of diabetes disease. To implement and evaluate this research, various steps have been performed as shown in figure 4.1. The first step is to understand and prepare the data set from UNRWA clinics, the second step is to implement data mining methods to build a classifier for extracting useful rules from the data set, and the third step is to evaluate the data mining system. In the fourth step, develop a mobile application and use the extracted rules from data mining system to predict the complications for diabetes patients based on data that affect diabetes. The fifth step is to evaluate the mobile application.

Figure 4.1: The steps of approach about using data mining and mobile application for patients of diabetes disease

This chapter is organized as follows. In section 4.1, presents the steps that taken to understand and prepare the data set for implementing data mining methods. Section 4.2 describes the data mining methods used to build a classifier for extracting useful rules from the data set. Next, discusses evaluation the data mining system (Section 4.3), describe steps to develop mobile application (Section 4.4), and
evaluating the mobile application (Section 4.5). Finally, chapter summary in section 4.6.

4.1. Collect and preparing the data set

The data of diabetic patients were obtained from the UNRWA database for the year 2015 in one excel file for each clinic of the 5 clinics as shown in the table 4.1.

4.1.1. Data set attributes:

Table 4.1: Diabetic patients’ instance counts

<table>
<thead>
<tr>
<th>#</th>
<th>Clinic Name</th>
<th>No. of instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Al Remal</td>
<td>639</td>
</tr>
<tr>
<td>2</td>
<td>Al Saftawi</td>
<td>294</td>
</tr>
<tr>
<td>3</td>
<td>Al Nusairat</td>
<td>468</td>
</tr>
<tr>
<td>4</td>
<td>Khan Younis</td>
<td>509</td>
</tr>
<tr>
<td>5</td>
<td>Al Shabora</td>
<td>206</td>
</tr>
</tbody>
</table>

The file contains patient’s personal data, health data, and blood glucose measurements. The data attributes describes in table 4.2:

Table 4.2: Diabetic patient’s data set attributes

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patient_ID</td>
<td>The patient number</td>
</tr>
<tr>
<td>2</td>
<td>PatientDOB</td>
<td>Patient Date of birth</td>
</tr>
<tr>
<td>3</td>
<td>PatientGender</td>
<td>Patient Gender (Male, Female)</td>
</tr>
<tr>
<td>4</td>
<td>Blood_Group</td>
<td>Patient Blood Group (A-, A+, AB-, AB+ … etc.)</td>
</tr>
</tbody>
</table>
| 5  | Age            | One of the risk factors,                 
|     |                | If patient age <= 45 years, Age risk factor value = 0                       |
|     |                | If patient age 46 – 55 years, Age risk factor value = 1                      |
|     |                | If patient age 56 – 64 years, Age risk factor value = 2                      |
|     |                | If patient age >= 65 years, Age risk factor value = 3 (see Appendix A)       |
| 6  | EBP            | Elevate blood pressure                                                       |
|     |                | One of the risk factors,                 
|     |                | If blood pressure controlled with lifestyle modification, EBP risk factor value = 0 |
|     |                | If blood pressure controlled with medications, EBP risk factor value = 1     |
|     |                | If blood pressure Uncontrolled < 160/100, EBP risk factor value = 2          |
|     |                | If blood pressure Uncontrolled >= 160/100, EBP risk factor value = 3 (see Appendix A) |
| 7  | DM             | Diabetes Mellitus                                                            |

One of the risk factors,
<table>
<thead>
<tr>
<th>#</th>
<th>Attribute Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Attribute Name</td>
<td>Description</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Diabetes Mellitus Controlled with lifestyle modification, DM risk factor value = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Diabetes Mellitus Controlled with medications, DM risk factor value = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Diabetes Mellitus Uncontrolled with medications, DM risk factor value = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Diabetes Mellitus Diabetes Mellitus with Proteinuria, DM risk factor value = 3 (see Appendix A)</td>
</tr>
<tr>
<td>8</td>
<td>Score_Smoking</td>
<td>One of the risk factors, No, Score_Smoking risk factor value = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes, Score_Smoking risk factor value = 3 (see Appendix A)</td>
</tr>
<tr>
<td>9</td>
<td>LD</td>
<td>Lipids disorders (Cholesterol)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>One of the risk factors, If Cholesterol &lt; 160 mg/dl, LD risk factor value = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Cholesterol 160 - 199 mg/dl, LD risk factor value = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Cholesterol 200 - 249 mg/dl, LD risk factor value = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Cholesterol ≥ 250 mg/dl, LD risk factor value = 3 (see Appendix A)</td>
</tr>
<tr>
<td>10</td>
<td>Score_BMI</td>
<td>Body Mass Index (كتلة الجسم)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>One of the risk factors, If BMI ≤ 29, Score_BMI risk factor value = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If BMI 30 – 34, Score_BMI risk factor value = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If BMI ≥ 35, Score_BMI risk factor value = 2 (see Appendix A)</td>
</tr>
<tr>
<td>11</td>
<td>Inactivity</td>
<td>Physical Activity (20-30 minutes of activities (e.g. walking) 3 minutes per week)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>One of the risk factors, Active, Inactivity risk factor value = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sedentary life, Inactivity risk factor value = 2 (see Appendix A)</td>
</tr>
<tr>
<td>12</td>
<td>Family_History</td>
<td>Family History (Close relatives with hereditary link (parents, brothers, sisters, uncles and aunts) with cardiovascular diseases including strokes, myocardial infarction (MI) and hypertension)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>One of the risk factors, Negative, Family_History risk factor value = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positive, Family_History risk factor value = 2 (see Appendix A)</td>
</tr>
<tr>
<td>13</td>
<td>Result1</td>
<td>blood glucose measurement</td>
</tr>
</tbody>
</table>

Patients’ data also contain attributes each of them is a class label of our experiment, which will be predicted by the classifier. These attributes explains the types of complications that a diabetic may have and determine whether he or she has these complications (Table 4.3). The value of 0 means the patient do not have this complication, while the value of 1 means the patient have this complication.
Table 4. 3: Diabetic complication attributes

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MI</td>
<td>Myocardial Infarction (Complication)</td>
</tr>
<tr>
<td>2</td>
<td>CHF</td>
<td>Congestive Heart Failure (Complication)</td>
</tr>
<tr>
<td>3</td>
<td>Stroke</td>
<td>Stroke (Complication)</td>
</tr>
<tr>
<td>4</td>
<td>ESRF</td>
<td>End-Stage Renal Failure (Complication)</td>
</tr>
<tr>
<td>5</td>
<td>Blindess</td>
<td>Blindness (Complication)</td>
</tr>
<tr>
<td>6</td>
<td>Ampution</td>
<td>Amputation (Complication)</td>
</tr>
</tbody>
</table>

A new aggregation attribute named “RiskFactor” has been creates to represents the summation of all risk factors values that lead to serious complications of diabetics. These risk factors attributes are “Age”, Elevate blood pressure “EBP”, Diabetes Mellitus “DM”, “Score_Smoking”, Lipids disorders (Cholesterol) “LD”, Body Mass Index “Score_BMI”, “Inactivity” and “Family_History”.

4.1.2. Data preprocessing

Data preprocessing for data mining is the set of techniques used prior to the application of a data mining method (García et al., 2015). Real-world data is often imperfect, containing inconsistencies and redundancies, these data are not directly applicable for a starting a data mining process. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing (Bramer; Jiawei Han, 2012).

4.1.2.1. Missing values:

Missing data are introduce due to various reasons, such as the data was not available or manual data entry procedures. Removing or filling missing data is an important operation to ensure that acceptable results are given. Removing rows is usually done when many attributes are missing from the row.

In our experiment, we replace a missing value with 0 as a replacement value for the following attributes (“MI”, “CHF”, “Stroke”, “ESRF”, “Blindess”, “Ampution”, “EBP”, “DM”, “Score_Smoking”, “LD”, “Score_BMI”, “Inactivity”, “Family_History”, “Result1”).

On the other hand, the missing values of patient’s age risk factor was derived according to the ranges shown in table 4.1 above, by calculating the age based on the patient’s date of birth.
4.1.2.2. Data Reduction:

Data reduction techniques can be applied to obtain a reduced representation of the data set, attribute subset selection is one of data reduction techniques. Attribute subset selection aim at selecting a subset of the attributes or features, which describe the data in order to obtain a more essential and compact representation of the available information.

In our experiment, we choose the attributes “PatientGender”, “Age”, “EBP”, “DM”, “Score_Smoking”, “LD”, “Score_BMI”, “Inactivity”, “Family_History” and “Result1”, because these attributes used for building a classifier to predict diabetes complications.

4.1.2.3. Data Transformation:

In data transformation, the data are transformed or consolidated into forms appropriate for mining, data discretization is a form of data transformation. Discretization is replace a numeric attribute values (e.g., age) by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior).

In our experiment, we discretized the following attributes:

- “PatientDOB” (Patient date of birth): after represents the date values as the year the patients were born:
- “Result1” (blood glucose measurement):
  - Low: With upper limit 110.
  - Normal: With upper limit 130.
  - High: With upper limit 1000.
- “RiskFactor” (the new aggregation attribute), this attribute discretized into three classes were classified according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A):
  - Low Risk: With upper limit 5.
  - Moderate Risk: With upper limit 9.

4.2. Data mining methods

Data mining methods are the main part of our approach in this research, it’s used to predict diabetic complications based on several risk factors such as patient age, lifestyle, smoking, blood glucose measurements and others. In this research, we choose classification method because it analysis diabetic patients data and extracts a model to clarify the set of cumulative risk factors that lead to serious complications of diabetics.
Below we will explain the steps that have been taken to prepare the data and create the classifier for each late complication (target class attributes) that may infect the patient by applying a decision tree classification method.

4.2.1. Myocardial Infarction (MI):

The following process (Figure 4.2) is the main process of the decision tree method that we applied on the dataset to build a classifier used to predict if the patient will be infect with myocardial infarction (MI), this method is implemented via RapidMiner tool:

![Figure 4.2: The main process of decision tree method in RapidMiner tool for Myocardial Infarction (MI) complication](image)

The previous figure 4.2 includes the following steps:

1- **Retrieve**: this step retrieves the dataset from data repository after it reads from the excel file which consists of 448 instances belongs to patients who are infected with Myocardial Infarction (MI).

2- **Replace Missing Values**: replaces missing values of selected attributes by a specified replacement (GmbH). In our experiment, we replace a missing value with 0 as a replacement value for the following attributes (“MI”, “CHF”, “Stroke”, “ESRF”, “Blindess”, “Amputation”, “Age”, “EBP”, “DM”, “Score_Smoking”, “LD”, “Score_BMI”, “Inactivity”, “Family_History”, “Result1”).
3- **Date to Numerical:** This operator changes the type of the selected date type attribute to a numeric type (GmbH). In our experiment, we change “PatientDOB” date attribute to represent as the year the patient was born.

4- **Discretize by User Specification:** This operator discretizes the selected numerical attributes into user specified classes. The user can define the classes by specifying the upper limits of each class (GmbH). We discretized the attribute “PatientDOB” (after using a date to numerical operator to represent the date values as the year the patients were born) into three classes:

5- **Filter Examples:** This Operator selects which instances of a dataset will be kept according to the matching of the given condition defined by user, and which instances will be removed (GmbH). In our experiment, we want to keep all instances with a value greater than 0 for “Result1” (blood glucose measurement) attribute.

6- **Discretize by User Specification:** We discretized the attribute “Result1” (blood glucose measurement) into three classes:
   - Low: With upper limit 110.
   - Normal: With upper limit 130.
   - High: With upper limit 1000.

7- **Generate Aggregation:** This operator generates a new attribute by performing the specified aggregation function for every instance of the selected attributes (GmbH). We used this operation to generate a new aggregation attribute named “RiskFactor” to represent the summation of all risk factors values that lead to serious complications of diabetics. These risk factors attributes are “Age”, “EBP”, “DM”, “Score_Smoking”, “LD”, “Score_BMI”, “Inactivity” and “Family_History”.

8- **Discretize by User Specification:** We discretized the new aggregation attribute “RiskFactor” into three classes, these classes were classified according to a medical bulletin named “Risk Scoring Protocol for Patients with Diabetes and/or Hypertension” (see Appendix A), obtained from the health center:
   - Low Risk: With upper limit 5.
   - Moderate Risk: With upper limit 9.

9- **Select Attributes:** This Operator selects a subset of attributes of a dataset and removes the others (GmbH). Because this process used for building a classifier to predict if the patient will be infect with myocardial infarction, we
choose the attributes “PatientGender”, “Age”, “EBP”, “DM”, “Score_Smoking”, “LD”, “Score_BMI”, “Inactivity”, “Family_History”, “Result1” and “MI”, and leave the attributes “CHF”, “Stroke”, “ESRF”, “Blindness” and “Amputation”, which refers to other complications.

10- **Numerical to Binominal**: This operator changes the type of the selected numeric attributes to a binominal type. This operator also maps all values of these attributes to corresponding binominal values (GmbH). In this process we choose the “MI” attribute to change its numeric type values to binominal type values, because it’s the predictable attribute in this process.

11- **Set Role**: This operator is used to change the role of one or more attributes, all attributes in the dataset is a regular attributes by default. when we changes an attribute to any other role, its called a special attribute (GmbH). In our process we changed the target role of the attribute “MI” to “label” as shown in (Figure 4.3), which means it’s a class label of our experiment.

![Figure 4.3: Change “MI” attribute to a class label](image)

12- **Split Validation**: This operator randomly splits up the dataset into a training set and test set and evaluates the model. This operator performs a split validation in order to estimate the accuracy of the model (GmbH). In our experiment, we choose a split ratio equal 0.7 as shown in (Figure 4.4), means that 70% of the dataset instances are used as a training set and the other 30% of instances are used as a test set.

![Figure 4.4: Split ratio](image)

13- **Decision Tree**: This Operator generates a decision tree model, which is used for classification.

14- **Apply Model**: This operator applies a model on a dataset. While the decision tree operator used for generates the classification model based on the training set, the apply model get a prediction on the test set (GmbH) (Figure 4.5).

15- **Performance Classification**: This operator is used for statistical performance evaluation of classification tasks (GmbH). in our experiment, we measure performance accuracy for the classification model.
4.2.2. Congestive Heart Failure (CHF):

The dataset consists of 401 instances belongs to patients who are infected with Congestive Heart Failure (CHF). The following process (Figure 4.6) is the main process of the decision tree method that we applied on the dataset to build a classifier used to predict if the patient will be infect with Congestive Heart Failure (CHF), this method is implemented via RapidMiner tool:

![Figure 4.6: The main process of decision tree method in RapidMiner tool for Congestive Heart Failure (CHF) complication](image)

The previous figure includes the same steps were done as in the decision tree method of Myocardial Infarction (MI) classifier except that the attribute “CHF” was selected as a class label in the Set Role operator (Figure 4.7).

![Figure 4.7: Change “CHF” attribute to a class label](image)
4.2.3. Stroke:

The dataset consists of 209 instances belongs to patients who are infected with Stroke. The following process (Figure 4.8) is the main process of the decision tree method that we applied on the dataset to build a classifier used to predict if the patient will be infected with Stroke, this method is implemented via RapidMiner tool:

![Figure 4.8: The main process of decision tree method in RapidMiner tool for Stroke complication](image)

The previous figure includes the same steps were done as in the decision tree method of Myocardial Infarction (MI) classifier except that the attribute “Stroke” was selected as a class label in the Set Role operator (Figure 4.9).

![Figure 4.9: Change “Stroke” attribute to a class label](image)

4.2.4. End Stage Renal Failure (ESRF):

The dataset consists of 652 instances belongs to patients who are infected with End Stage Renal Failure (ESRF). The following process (Figure 4.10) is the main process of the decision tree method that we applied on the dataset to build a classifier used to predict if the patient will be infected with End Stage Renal Failure (ESRF), this method is implemented via RapidMiner tool:
The previous figure includes the same steps were done as in the decision tree method of Myocardial Infarction (MI) classifier except that the attribute “ESRF” was selected as a class label in the Set Role operator (Figure 4.11).

Figure 4. 11: Change “ESRF” attribute to a class label

4.2.5. Blindness:

The dataset consists of 314 instances belongs to patients who are infected with Blindness. The following process (Figure 4.12) is the main process of the decision tree method that we applied on the dataset to build a classifier used to predict if the patient will be infect with Blindness, this method is implemented via RapidMiner tool:
Figure 4.12: The main process of decision tree method in RapidMiner tool for Blindness complication

The previous figure includes the same steps were done as in the decision tree method of Myocardial Infarction (MI) classifier except that the attribute “Blindness” was selected as a class label in the Set Role operator (Figure 4.13).

Figure 4.13: Change “Blindness” attribute to a class label

4.2.6. Amputation:

The dataset consists of 305 instances belongs to patients who are infected with Amputation. The following process (Figure 4.14) is the main process of the decision tree method that we applied on the dataset to build a classifier used to predict if the patient will be infect with Amputation, this method is implemented via RapidMiner tool:
Figure 4.14: The main process of decision tree method in RapidMiner tool for Amputation complication

The previous figure includes the same steps were done as in the decision tree method of Myocardial Infarction (MI) classifier except that the attribute “Amputation” was selected as a class label in the Set Role operator (Figure 4.15).

Figure 4.15: Change “Amputation” attribute to a class label

4.3. Evaluate the data mining system

Data mining classification is a two-step process. Learning step where a classification model (classifier) is constructed based on previous data and classification step where the model is uses to classify new data when the model’s accuracy is acceptable.

First, the data set separated into training set and test set. In the learning step, most of data set used as a training set including the associated class label attribute which its value represents a category or class for each data instance. In the classification step, a small portion and independent of the training set of data is used as a test set to estimate the predictive accuracy of the classifier. The accuracy of a classifier on a given test set is the percentage of test set instances that are correctly classified by the classifier. The class label of each test instance is compared with the
learned classifier’s class prediction for that instance. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify new data instances for which the class label is not known (Oded Maimon, 2010).

The most obvious criterion to use for estimating the performance of a classifier is predictive accuracy. And where it is not possible to establish the predictive accuracy on a dataset that used to generate the classifier, the available data is split into two parts called a training set and a test set (Figure 4.16). First, the training set is used to construct a classifier. The classifier is then used to predict the classification for the instances in the test set.

![Figure 4.16: Training set and test set (Bramer)](image)

There are four additional terms used in computing evaluation measures:

- **True positives (TP):** refer to the positive instances that correctly labeled by the classifier. Let TP be the number of true positives.

- **True negatives (TN):** refer to the negative instances that correctly labeled by the classifier. Let TN be the number of true negatives.

- **False positives (FP):** refer to the negative instances that incorrectly labeled as positive. (e.g., instances of class infection of blindness = no for which the classifier predicted infection of blindness = yes). Let FP be the number of false positives.

- **False negatives (FN):** refer to the positive instances that mislabeled as negative. (e.g., instances of class infection of blindness = yes for which the classifier predicted infection of blindness = no). Let FN be the number of false negatives.

These terms are uses in the confusion matrix (Figure 4.17). A confusion matrix is a table used to describe the performance of a classification by the classifier on a set of test data for which the true values are known. TP and TN tell us when the
classifier is getting things right, while FP and FN tell us when the classifier is getting things wrong (Jiawei Han, 2012; Powers, 2011).

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>P</td>
<td>TP (True Positives)</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>FN (False Negatives)</td>
</tr>
<tr>
<td>N</td>
<td>P</td>
<td>FP (False Positives)</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>TN (True Negatives)</td>
</tr>
</tbody>
</table>

**Figure 4.17:** Confusion matrix

The accuracy of a classifier on any given test set is the percentage of test set instances that are correctly classified by the classifier (Jiawei Han, 2012; Sammut & Webb, 2011).

\[
\text{Accuracy (AC)} = \frac{TP + TN}{TP + FN + FP + TN} \tag{4.1}
\]

Another measure to evaluate the classifier is a precision-recall measure. Precision is the fraction of retrieved instances by the classifier that are relevant, while recall is the fraction of relevant instances that are retrieved by the classifier (Jiawei Han, 2012).

\[
\text{precision} = \frac{TP}{TP + FP} \tag{4.2}
\]

\[
\text{recall} = \frac{TP}{TP + FN} \tag{4.3}
\]

To calculate the harmonic mean between precision and recall, an F-score measure is used. F-score reaches its best value at 1 and worst at 0 (Jiawei Han, 2012; Wikipedia).

\[
F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4.4}
\]
4.4. Develop mobile application

Mobile applications come in two formats, native applications and mobile web applications. In our research we used a native applications, which build specifically for each mobile platform and installed on the device itself.

To implement the rules that extracted from the classification model, we used a rule extraction Java API routines to create six Java classes, one for each classifier. The output result for each class indicates whether the patient will infect with that complication or not, based on the data entered.

As shown in Figures 4.18 to 4.22 below, the interface of the mobile application is in Arabic, since some patients may not understand English:

Application inputs:

- **Patient Gender (الجنس):** The User can choose either male or female. This choice refers to “PatientGender” attribute.

- **Length (الطول) and Weight (الوزن):** The user must insert the patient's length in centimeter unit and weight in kilograms. These values are used to calculate the patient’s Body Mass Index (BMI). BMI is calculated by the equation
kg/m² where kg is a person’s weight in kilograms and m² is their height in metres squared (Wikipedia). The calculated BMI Value refers to “Score_BMI” attribute.

- **Blood glucose measurement** (قياس السكر في الدم): The user should choose the appropriate answer according to the blood glucose measurement. This choice refers to “Result1” attribute.

  Figure 4.19: Mobile application screenshot 2

- **Age** (العمر): The user should choose the appropriate answer according to patient’s age. This choice refers to “Age” attribute.

- **Blood Pressure** (ضغط الدم): The user should choose the appropriate answer according to the blood pressure measurement. This choice refers to “EBP” attribute.
Figure 4. 20: Mobile application screenshot 3

- **Diabetes Mellitus (حالة مرض السكر):** The user should choose the appropriate answer according to patient’s diabetes mellitus status. This choice refers to “DM” attribute.

- **Smoking (التدخين من أي نوع):** The user can choose either yes or no. This choice refers to “Score_Smoking” attribute.

- **Lipids disorders (Cholesterol) (قياس الكولسترول في الدم):** The user should choose the appropriate answer according to patient’s cholestrol measurement. This choice refers to “LD” attribute.
- **Physical Activity** (النشاط البدني): The user can choose either active (فعال) or sedentary life (حياة خاملة). This choice refers to “Inactivity” attribute.
- **Family History** (هل صيب أحد الأقارب من الدرجة الأولى بمرض السكري أو أمراض القلب والالأوعية الدموية): The user can choose either yes or no. This choice refers to “Family_History” attribute.
Application Outputs:

After completion of the process of insertion and selection of data and factors affecting the symptoms of diabetes, the user can see the expected complications according to the rules that extracted from the data mining system. The complications that may infect the patient are:

- **Amputation** (بتر أطراف): This value refers to “Amputation” attribute.
- **Blindness** (تلف شبكية العين): This value refers to “Blindness” attribute.
- **Congestive Heart Failure** (هبوط القلب المزمن): This value refers to “CHF” attribute.
- **End Stage Renal Failure** (الفشل الكلوي الكامن): This value refers to “ESRF” attribute.
- **Myocardial Infarction** (الجثحة الصردية): This value refers to “MI” attribute.
- **Stroke** (الجلطات الدماغية): This value refers to “Stroke” attribute.
4.5. Evaluate the mobile application

To evaluate the mobile application, we conducted a questionnaire to measure the acceptance of experts and patients to the idea of using mobile applications that serve as indicators for the emergence of potential complications of chronic diseases and diabetes in particular, in addition to measuring the ease of use of the application.

The questionnaire was distributed to 43 respondents. It consists of two sections; the first section contains the personal data of the respondent who participates in the questionnaire (doctor or patient, age group and educational level). The second section consists of seven questions (Table 4.3) with a scale of response Strongly Agree, Agree, Neutral, Disagree and Strongly Disagree.

<table>
<thead>
<tr>
<th>Table 4.4: The mobile application questionnaire questions</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>هل من المهم استخدام تكنولوجيا المعلومات لمعرفة المضاعفات المحتملة لمرض السكري؟</td>
<td>1</td>
</tr>
<tr>
<td>هل تطبيق المحمول سهل الاستخدام؟</td>
<td>2</td>
</tr>
<tr>
<td>هل تطبيق المحمول سهل التعلم؟</td>
<td>3</td>
</tr>
<tr>
<td>هل استفدت من استخدام التطبيق لمعرفة المضاعفات المحتملة لمرض السكري؟</td>
<td>4</td>
</tr>
<tr>
<td>هل فهمت النتائج التي أعطتها التطبيق؟</td>
<td>5</td>
</tr>
<tr>
<td>هل تتصفح الآخرين لاستخدام التطبيق لمعرفة مضاعفات مرض السكري؟</td>
<td>6</td>
</tr>
<tr>
<td>هل ترغب في رؤية تكنولوجيا مماثلة لتشخيص الأمراض المزمنة الأخرى؟</td>
<td>7</td>
</tr>
</tbody>
</table>

To measure respondent’s acceptability for each question, we put a weight for each response so that the weight of the answer "Strongly Disagree" equals 0, "Disagree" equals 25, "Neutral" equals 50, "Agree" equals 75 and the weight of response "Strongly agree" equals 100.

For example, to calculate the acceptance ratio (AR) for the 6th question:

\[ AR(6) = \frac{(No\ of\ Strongly\ Disagree \times \text{response\ weight}) + (No\ of\ Disagree \times \text{response\ weight}) + (No\ of\ Neutral \times \text{response\ weight}) + (No\ of\ Agree \times \text{response\ weight}) + (No\ of\ Strongly\ Agree \times \text{response\ weight})}{\text{Total No. of respondents}} \]
Then, $AR(6) = ((0 \times 0) +$ \\
$(0 \times 25) +$ \\
$(3 \times 50) +$ \\
$(16 \times 75) +$ \\
$(24 \times 100)) / 43$ \\
$= 87.21\%$

Finally, to calculate the overall acceptance ratio for all questions, we calculate the average of all question’s acceptance ratios.

### 4.6. Summary

In this chapter, we presented our experiments; we discussed the dataset attributes and described the data preprocessing techniques that used prior to the application of the data mining method. In addition, we presented the steps that have been performed to implement and evaluate the data mining system and the mobile application as well.
Chapter 5
Experiments and Results
In this chapter, we present and discuss the experiments and the results of our research. Section 5.1, presents the experiments setup that includes experimental environment and tools, Section 5.2 presents the results of the data mining model, section 5.3 presents the evaluation results of the data mining model. Section 5.4 presents the evaluation results of the mobile application questionnaire. Finally, chapter summary in section 5.5.

5.1. Experiments Setup:

In this section, we will describe the experimental environment, and determine the tools that were used in the experiments. Finally, determine the setting of the experiments in the research.

5.1.1. Experimental Environment:

We applied experiments on a machine with properties that is Intel(R) Core(TM) i7-2670QM CPU @ 2.20GHz, 8.00 GB RAM, 500 GB hard disk drive and Windows 10, 64-bit operating system installed.

5.1.2. Experimental Tools:

1- RapidMiner program: A software platform used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results visualization, validation and optimization.

2- Easy Rule Java API: It is a Java “rules engine”; we used it in our experiments to apply the results out from data mining system in the mobile application (see Appendix B for a sample code).

3- Android Studio: The official IDE for Android application development.

4- Microsoft Office Excel: Used to organize and store datasets in tables, do some simple preprocessing.

5.2. Results of the data mining model

In this section, we will present the results of the data mining model for each diabetic complications:

5.2.1. Myocardial Infarction (MI)

Figure 5.1 and Table 5.1 represents the rules extracted from the experiment dataset after applying decision tree classification method on 115 instances, which illustrate the factors leading to the probability of infection of myocardial infarction (MI).
Figure 5.1: Myocardial Infarction (MI) decision tree output result

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>if(RiskFactor = High) then the probability of infection of Myocardial Infarction (MI) = True</td>
</tr>
<tr>
<td>Rule 2</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = High) and (LD(Lipids disorders (Cholesterol)) &gt; 0)) then the probability of infection of Myocardial Infarction (MI) = True</td>
</tr>
<tr>
<td>Rule 3</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = High) and (LD(Lipids disorders (Cholesterol)) &lt;= 0)) then the probability of infection of Myocardial Infarction (MI) = False</td>
</tr>
<tr>
<td>Rule 4</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (EBP(Elevate blood pressure) &gt; 0)) then the probability of infection of Myocardial Infarction (MI) = True</td>
</tr>
<tr>
<td>Rule 5</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (EBP(Elevate blood pressure) &lt;= 0) and (PatientGender = Female) and (Age &gt; 2)) then</td>
</tr>
<tr>
<td>Rule #</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Rule 6</td>
<td>If((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (EBP(Elevate blood pressure) &lt;= 0) and (PatientGender = Female) and (Age &lt;= 2)) then the probability of infection of Myocardial Infarction (MI) = False</td>
</tr>
<tr>
<td>Rule 7</td>
<td>If((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (EBP(Elevate blood pressure) &lt;= 0) and (PatientGender = Female) and (Age &lt;= 2)) then the probability of infection of Myocardial Infarction (MI) = True</td>
</tr>
<tr>
<td>Rule 8</td>
<td>If((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (EBP(Elevate blood pressure) &lt;= 0) and (PatientGender = Male) and (Age &gt; 1)) then the probability of infection of Myocardial Infarction (MI) = True</td>
</tr>
<tr>
<td>Rule 9</td>
<td>If((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (EBP(Elevate blood pressure) &lt;= 0) and (PatientGender = Male) and (Age &gt; 1)) then the probability of infection of Myocardial Infarction (MI) = False</td>
</tr>
<tr>
<td>Rule 10</td>
<td>If((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (EBP(Elevate blood pressure) &lt;= 0)) and</td>
</tr>
<tr>
<td>Rule #</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td></td>
<td>(PatientGender = Male) and (Age &lt;= 1) then the probability of infection of Myocardial Infarction (MI) = False</td>
</tr>
<tr>
<td>Rule 11</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = Normal)) then the probability of infection of Myocardial Infarction (MI) = False</td>
</tr>
<tr>
<td>Rule 12</td>
<td>if(RiskFactor = Moderate) then the probability of infection of Myocardial Infarction (MI) = True</td>
</tr>
</tbody>
</table>

From the rules extracted above, patients are likely to be exposed to myocardial infarction (MI) in the following conditions:

- High or moderate risk factors (Rule 1, Rule 12), which means that the risk score is greater than 6 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A).
- A blood glucose measurement greater than 130, with a lipids disorders (Cholesterol) greater than 159 mg/dl (Rule 2).
- An elevate blood pressure controlled by medications or uncontrolled at all, with a blood glucose measurement less than 111 (Rule 4).

### 5.2.2. Congestive Heart Failure (CHF)

Figure 5.2 and table 5.2 represents the rules extracted from the experiment dataset after applying decision tree classification method on 111 instances, which illustrate the factors leading to the probability of infection of congestive heart failure (CHF).
**Figure 5.2**: Congestive Heart Failure (CHF) decision tree output result

**Table 5.2**: Congestive Heart Failure (CHF) rules description

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>if((RiskFactor = High) then the probability of infection of Congestive Heart Failure (CHF) = True</td>
</tr>
<tr>
<td>Rule 2</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = High) and (EBP(Elevate blood pressure) &gt; 0)) then the probability of infection of Congestive Heart Failure (CHF) = True</td>
</tr>
<tr>
<td>Rule 3</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = High) and (EBP(Elevate blood pressure) &lt;= 0)) then the probability of infection of Congestive Heart Failure (CHF) = False</td>
</tr>
<tr>
<td>Rule 4</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (Age &lt;= 2)) then the probability of infection of Congestive Heart Failure (CHF) = False</td>
</tr>
<tr>
<td>Rule 5</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (Age &gt; 2) and (PatientGender = Female)) then the probability of infection of Congestive Heart Failure (CHF) = False</td>
</tr>
<tr>
<td>Rule #</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Rule 6</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (Age &gt; 2) and (PatientGender = Male)) then the probability of infection of Congestive Heart Failure (CHF) = True</td>
</tr>
<tr>
<td>Rule 7</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = Normal)) then the probability of infection of Congestive Heart Failure (CHF) = False</td>
</tr>
<tr>
<td>Rule 8</td>
<td>if(RiskFactor = Moderate) then the probability of infection of Congestive Heart Failure (CHF) = True</td>
</tr>
</tbody>
</table>

From the rules extracted above, patients are likely to be exposed to congestive heart failure (CHF) in the following conditions:

- High or moderate risk factors (Rule 1, Rule 8), which means that the risk score is greater than 6 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A).
- A blood glucose measurement greater than 130, with an elevate blood pressure controlled by medications or uncontrolled at all (Rule 2).
- Males greater than 65 years old (Rule 6).
5.2.3. Stroke

Figure 5.3 and table 5.3 represents the rules extracted from the experiment dataset after applying decision tree classification method on 56 instances, which illustrate the factors leading to the probability of infection of stroke.

**Figure 5.3:** Stroke decision tree output result

**Table 5.3:** Stroke rules description

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
</table>
| Rule 1 | if(RiskFactor = High)then  
 the probability of infection of Stroke = True |
| Rule 2 | if((RiskFactor = Low)and  
 (Result1(blood glucose measurement) = High)and  
 (Age > 0)and  
 (Age > 1)and  
 (PatientGender = Female))then  
 the probability of infection of Stroke = False |
| Rule 3 | if((RiskFactor = Low)and  
 (Result1(blood glucose measurement) = High)and  
 (Age > 0)and  
 (Age > 1)and  
 (PatientGender = Male)and  
 (Age > 2))then |
<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 4</td>
<td>$\text{if}((\text{RiskFactor} = \text{Low}) \text{and} (\text{Result1(blood glucose measurement)} = \text{High}) \text{and} (\text{Age} &gt; 0) \text{and} (\text{Age} &gt; 1) \text{and} (\text{PatientGender} = \text{Male}) \text{and} (\text{Age} \leq 2)) \text{then}$ the probability of infection of Stroke = False</td>
</tr>
<tr>
<td>Rule 5</td>
<td>$\text{if}((\text{RiskFactor} = \text{Low}) \text{and} (\text{Result1(blood glucose measurement)} = \text{High}) \text{and} (\text{Age} &gt; 0) \text{and} (\text{Age} \leq 1)) \text{then}$ the probability of infection of Stroke = False</td>
</tr>
<tr>
<td>Rule 6</td>
<td>$\text{if}((\text{RiskFactor} = \text{Low}) \text{and} (\text{Result1(blood glucose measurement)} = \text{High}) \text{and} (\text{Age} \leq 0)) \text{then}$ the probability of infection of Stroke = False</td>
</tr>
<tr>
<td>Rule 7</td>
<td>$\text{if}((\text{RiskFactor} = \text{Low}) \text{and} (\text{Result1(blood glucose measurement)} = \text{Low}) \text{and} (\text{EBP(Elevate blood pressure)} &gt; 0)) \text{then}$ the probability of infection of Stroke = True</td>
</tr>
<tr>
<td>Rule 8</td>
<td>$\text{if}((\text{RiskFactor} = \text{Low}) \text{and} (\text{Result1(blood glucose measurement)} = \text{Low}) \text{and} (\text{EBP(Elevate blood pressure)} \leq 0) \text{and} (\text{Age} &gt; 0) \text{and} (\text{Age} &gt; 1) \text{and} (\text{PatientGender} = \text{Female})) \text{then}$ the probability of infection of Stroke = False</td>
</tr>
</tbody>
</table>
| Rule 9 | $\text{if}((\text{RiskFactor} = \text{Low}) \text{and} (\text{Result1(blood glucose measurement)} = \text{Low}) \text{and} (\text{EBP(Elevate blood pressure)} \leq 0) \text{and} (\text{Age} > 0) \text{and}$}
From the rules extracted above, patients are likely to be exposed to stroke in the following conditions:

- High or moderate risk factors (Rule 1, Rule 13), which means that the risk score is greater than 6 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A).
- Males greater than 65 years old, with blood glucose measurement greater than 130 (Rule 3).
- An elevate blood pressure are controlled by medications or uncontrolled at all but blood glucose measurement less than 111 (Rule 7).
- Males greater than 55 years old, but blood glucose measurement less than 111 and an elevate blood pressure are controlled with lifestyle modification (Rule 9).
5.2.4. End Stage Renal Failure (ESRF)

Figure 5.4 and table 5.4 represents the rules extracted from the experiment dataset after applying decision tree classification method on 163 instances, which illustrate the factors leading to the probability of infection of end stage renal failure (ESRF).

![Decision Tree Diagram](image)

**Figure 5.4: End Stage Renal Failure (ESRF) decision tree output result**

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>if(RiskFactor = High) then the probability of infection of End Stage Renal Failure (ESRF) = True</td>
</tr>
<tr>
<td>Rule 2</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = High) and (EBP(Elevate blood pressure) &gt; 1)) then the probability of infection of End Stage Renal Failure (ESRF) = True</td>
</tr>
<tr>
<td>Rule 3</td>
<td>if((RiskFactor = Low) and (Result1(blood glucose measurement) = High) and (EBP(Elevate blood pressure) &lt;= 1) and (Age &gt; 2)) then the probability of infection of End Stage Renal Failure (ESRF) = False</td>
</tr>
<tr>
<td>Rule 4</td>
<td>if((RiskFactor = Low) and</td>
</tr>
<tr>
<td>Rule #</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td></td>
<td>((\text{Result1(blood glucose measurement)} = \text{High}) \land (\text{EBP(Elevate blood pressure)} \leq 1) \land (\text{Age} \leq 2) \land (\text{Age} \leq 0))) then (\text{the probability of infection of End Stage Renal Failure (ESRF)} = \text{False})</td>
</tr>
<tr>
<td>Rule 5</td>
<td>if((\text{RiskFactor} = \text{Low}) \land (\text{Result1(blood glucose measurement)} = \text{High}) \land (\text{EBP(Elevate blood pressure)} \leq 1) \land (\text{Age} \leq 2) \land (\text{Age} &gt; 0) \land (\text{PatientGender} = \text{Female})) then (\text{the probability of infection of End Stage Renal Failure (ESRF)} = \text{False})</td>
</tr>
<tr>
<td>Rule 6</td>
<td>if((\text{RiskFactor} = \text{Low}) \land (\text{Result1(blood glucose measurement)} = \text{High}) \land (\text{EBP(Elevate blood pressure)} \leq 1) \land (\text{Age} \leq 2) \land (\text{Age} &gt; 0) \land (\text{PatientGender} = \text{Male}) \land (\text{Age} &gt; 1)) then (\text{the probability of infection of End Stage Renal Failure (ESRF)} = \text{True})</td>
</tr>
<tr>
<td>Rule 7</td>
<td>if((\text{RiskFactor} = \text{Low}) \land (\text{Result1(blood glucose measurement)} = \text{High}) \land (\text{EBP(Elevate blood pressure)} \leq 1) \land (\text{Age} \leq 2) \land (\text{Age} &gt; 0) \land (\text{PatientGender} = \text{Male}) \land (\text{Age} \leq 1)) then (\text{the probability of infection of End Stage Renal Failure (ESRF)} = \text{False})</td>
</tr>
<tr>
<td>Rule 8</td>
<td>if((\text{RiskFactor} = \text{Low}) \land (\text{Result1(blood glucose measurement)} = \text{Low}) \land (\text{Age} &gt; 0)) then (\text{the probability of infection of End Stage Renal Failure (ESRF)} = \text{False})</td>
</tr>
<tr>
<td>Rule #</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>Rule 9</td>
<td>if((RiskFactor = Low)(\text{and})(\text{Result1(blood glucose measurement)} = \text{Low})(\text{and})(\text{Age &gt; 0})(\text{and})(\text{Age &lt;= 2})(\text{and})(\text{EBP(Elevate blood pressure)} &gt; 1))then the probability of infection of End Stage Renal Failure (ESRF) = True</td>
</tr>
<tr>
<td>Rule 10</td>
<td>if((RiskFactor = Low)(\text{and})(\text{Result1(blood glucose measurement)} = \text{Low})(\text{and})(\text{Age &gt; 0})(\text{and})(\text{Age &lt;= 2})(\text{and})(\text{EBP(Elevate blood pressure)} &lt;= 1)(\text{and})(PatientGender = Female))then the probability of infection of End Stage Renal Failure (ESRF) = True</td>
</tr>
<tr>
<td>Rule 11</td>
<td>if((RiskFactor = Low)(\text{and})(\text{Result1(blood glucose measurement)} = \text{Low})(\text{and})(\text{Age &gt; 0})(\text{and})(\text{Age &lt;= 2})(\text{and})(\text{EBP(Elevate blood pressure)} &lt;= 1)(\text{and})(PatientGender = Male))then the probability of infection of End Stage Renal Failure (ESRF) = False</td>
</tr>
<tr>
<td>Rule 12</td>
<td>if((RiskFactor = Low)(\text{and})(\text{Result1(blood glucose measurement)} = \text{Low})(\text{and})(Age &lt;= 0))then the probability of infection of End Stage Renal Failure (ESRF) = False</td>
</tr>
<tr>
<td>Rule 13</td>
<td>if((RiskFactor = Low)(\text{and})(Result1(blood glucose measurement) = Normal))then the probability of infection of End Stage Renal Failure (ESRF) = False</td>
</tr>
<tr>
<td>Rule 14</td>
<td>if(RiskFactor = Moderate)then the probability of infection of End Stage Renal Failure (ESRF) = True</td>
</tr>
</tbody>
</table>
From the rules extracted above, patients are likely to be exposed to end stage renal failure (ESRF) in the following conditions:

- High or moderate risk factors (Rule 1, Rule 14), which means that the risk score is greater than 6 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A).
- A blood glucose measurement greater than 130, with uncontrolled elevate blood pressure (Rule 2).
- Males less than 56 years old and a blood glucose measurement greater than 130 with controlled elevate blood pressure even by lifestyle modification or medications (Rule 6).
- Patients less than 65 years old with uncontrolled elevate blood pressure (Rule 9).
- Females less than 65 years old with controlled elevate blood pressure even by lifestyle modification or medications (Rule 10).

5.2.5. Blindness

Figure 5.5 and table 5.5 represents the rules extracted from the experiment dataset after applying decision tree classification method on 87 instances, which illustrate the factors leading to the probability of infection of blindness.

![Blindness decision tree output result](image)

**Table 5.5: Blindness rules description**

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>if (RiskFactor = High) then the probability of infection of Blindness = True</td>
</tr>
<tr>
<td>Rule 2</td>
<td>if ((RiskFactor = Low) and</td>
</tr>
<tr>
<td>Rule #</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td></td>
<td>(Result1(blood glucose measurement) = High))then</td>
</tr>
<tr>
<td></td>
<td>the probability of infection of Blindness = False</td>
</tr>
</tbody>
</table>

**Rule 3**

if((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (PatientGender = Female) and (Age > 2)) then  
the probability of infection of Blindness = False

**Rule 4**

if((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (PatientGender = Female) and (Age <= 2)) then  
the probability of infection of Blindness = True

**Rule 5**

if((RiskFactor = Low) and (Result1(blood glucose measurement) = Low) and (PatientGender = Male)) then  
the probability of infection of Blindness = False

**Rule 6**

if((RiskFactor = Low) and (Result1(blood glucose measurement) = Normal)) then  
the probability of infection of Blindness = False

**Rule 7**

if(RiskFactor = Moderate) then  
the probability of infection of Blindness = False

From the rules extracted above, patients are likely to be exposed to blindness in the following conditions:

- High risk factors (Rule 1), which means that the risk score is greater than 9 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A).
- Females less than 65 years old with blood glucose measurement less than 111 (Rule 4).
5.2.6. Amputation

Figure 5.6 and table 5.6 represents the rules extracted from the experiment dataset after applying decision tree classification method on 78 instances, which illustrate the factors leading to the probability of infection of amputation.

![Amputation decision tree output result](image)

**Figure 5.6:** Amputation decision tree output result

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>if(RiskFactor = High) \then \the probability of infection of Amputation = True</td>
</tr>
<tr>
<td>Rule 2</td>
<td>if((RiskFactor = Moderate)\and (\text{PatientGender} = \text{Male})) \and (\text{Result1(blood glucose measurement)} = \text{Low}) then \the probability of infection of Amputation = True</td>
</tr>
<tr>
<td>Rule 3</td>
<td>if((RiskFactor = Moderate)\and (\text{PatientGender} = \text{Male})) \and (\text{Result1(blood glucose measurement)} = \text{Normal}) then \the probability of infection of Amputation = False</td>
</tr>
<tr>
<td>Rule 4</td>
<td>if((RiskFactor = Moderate)\and (\text{PatientGender} = \text{Male})) \and (\text{Result1(blood glucose measurement)} = \text{High}) \and \text{DM(Diabetes Mellitus)} &gt; 1)) then</td>
</tr>
</tbody>
</table>

**Table 5.6:** Amputation rules description
<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
</table>
| Rule 5 | if((RiskFactor = Moderate)\textbf{and}  \\
|        | (PatientGender = Male)\textbf{and}  \\
|        | (Result1(\textit{blood glucose measurement}) = High)\textbf{and}  \\
|        | (\textit{DM}(Diabetes Mellitus) <= 1)\textbf{and}  \\
|        | (Family_History > 1))then  \\
|        | the probability of infection of Amputation = True |
| Rule 6 | if((RiskFactor = Moderate)\textbf{and}  \\
|        | (PatientGender = Male)\textbf{and}  \\
|        | (Result1(\textit{blood glucose measurement}) = High)\textbf{and}  \\
|        | (\textit{DM}(Diabetes Mellitus) <= 1)\textbf{and}  \\
|        | (Family_History <= 1))then  \\
|        | the probability of infection of Amputation = False |
| Rule 7 | if((RiskFactor = Low)\textbf{and}  \\
|        | (Result1(\textit{blood glucose measurement}) = High))then  \\
|        | the probability of infection of Amputation = False |
| Rule 8 | if((RiskFactor = Low)\textbf{and}  \\
|        | (Result1(\textit{blood glucose measurement}) = Normal))then  \\
|        | the probability of infection of Amputation = False |
| Rule 9 | if((RiskFactor = Low)\textbf{and}  \\
|        | (Result1(\textit{blood glucose measurement}) = Low)\textbf{and}  \\
|        | (PatientGender = Male))then  \\
|        | the probability of infection of Amputation = False |
| Rule 10| if((RiskFactor = Low)\textbf{and}  \\
|       | (Result1(\textit{blood glucose measurement}) = Low)\textbf{and}  \\
|       | (PatientGender = Female)\textbf{and}  \\
|       | (Age > 2))then  \\
|       | the probability of infection of Amputation = False |
| Rule 11| if((RiskFactor = Low)\textbf{and}  \\
|       | (Result1(\textit{blood glucose measurement}) = Low)\textbf{and}  \\
|       | (PatientGender = Female)\textbf{and}  \\
|       | (Age <= 2))then  \\
<p>|       | the probability of infection of Amputation = False |</p>
<table>
<thead>
<tr>
<th>Rule #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>the probability of infection of Amputation = True</td>
</tr>
</tbody>
</table>

From the rules extracted above, patients are likely to be exposed to amputation in the following conditions:

- High risk factors (Rule 1), which means that the risk score is greater than 9 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A).
- Males with blood glucose measurement less than 111 and moderate risk factors which means that the risk score is between 6 and 9 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A) (Rule 2).
- Males with blood glucose measurement greater than 130 and moderate risk factors which means that the risk score is between 6 and 9 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A), with a diabetes mellitus uncontrolled with medications or the patient have a diabetes mellitus with Proteinuria (Rule 4).
- Males with blood glucose measurement greater than 130 and moderate risk factors which means that the risk score is between 6 and 9 according to the Risk Scoring Protocol for Patients with Diabetes and/or Hypertension (see Appendix A), with a diabetes mellitus controlled even with lifestyle modification or medications and the patient has a close relatives with hereditary link (parents, brothers, sisters, uncles and aunts) with cardiovascular diseases including strokes, myocardial infarction (MI) and hypertension (Rule 5).
- Females less than 65 years old with blood glucose measurement less than 111 (Rule 11).

### 5.3. Evaluation results of the data mining model

In this section, we will present the evaluation results of the data mining model for diabetic complications by calculating the precision, recall, F-score and accuracy for each classifier using the confusion matrix. Table 5.7 below illustrates the evaluation results for each classifier.
### Table 5.7: Accuracy results of the data mining system

<table>
<thead>
<tr>
<th>#</th>
<th>Classifier</th>
<th>No of instances</th>
<th>True Positive (TP)</th>
<th>False Positive (FP)</th>
<th>False Negative (FN)</th>
<th>True Negative (TN)</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Myocardial Infarction (MI)</td>
<td>115</td>
<td>41</td>
<td>0</td>
<td>5</td>
<td>69</td>
<td>100.00%</td>
<td>98.13%</td>
<td>94.25%</td>
<td>95.65%</td>
</tr>
<tr>
<td>2</td>
<td>Congestive Heart Failure (CHF)</td>
<td>111</td>
<td>13</td>
<td>0</td>
<td>2</td>
<td>96</td>
<td>100.00%</td>
<td>86.67%</td>
<td>92.86%</td>
<td>98.20%</td>
</tr>
<tr>
<td>3</td>
<td>Stroke</td>
<td>56</td>
<td>23</td>
<td>0</td>
<td>3</td>
<td>30</td>
<td>100.00%</td>
<td>88.46%</td>
<td>93.88%</td>
<td>94.64%</td>
</tr>
<tr>
<td>4</td>
<td>End Stage Renal Failure (ESRF)</td>
<td>163</td>
<td>56</td>
<td>8</td>
<td>0</td>
<td>99</td>
<td>87.50%</td>
<td>100.00%</td>
<td>93.33%</td>
<td>95.09%</td>
</tr>
<tr>
<td>5</td>
<td>Blindness</td>
<td>87</td>
<td>51</td>
<td>0</td>
<td>1</td>
<td>35</td>
<td>100.00%</td>
<td>98.08%</td>
<td>99.03%</td>
<td>98.85%</td>
</tr>
<tr>
<td>6</td>
<td>Amputation</td>
<td>78</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>97.92%</strong></td>
<td><strong>93.72%</strong></td>
<td><strong>95.56%</strong></td>
<td><strong>97.07%</strong></td>
</tr>
</tbody>
</table>

For example, the classifier has predicted that 69 instances are not infected with myocardial infarction (MI) correctly \((\text{which refers to the term True Negatives (TN) in the confusion matrix})\), and failed with 5 instances \((\text{which refers to the term False Negatives (FN) in the confusion matrix})\).

In addition, the classifier has predicted that 41 instances are infected with myocardial infarction (MI) correctly \((\text{which refers to the term True Positives (TP) in the confusion matrix})\), and failed with 0 instances \((\text{which refers to the term False Positives (FP) in the confusion matrix})\).

The precision of the classifier is the percentage of retrieved instances by the classifier that are relevant, the precision percentage of myocardial infarction (MI) classifier is:

\[
\text{Precision (MI)} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]

\[
= \frac{41}{(41 + 0)}
\]

\[
= 100\%
\]

The recall of the classifier is the percentage of relevant instances that are retrieved by the classifier, the recall percentage of myocardial infarction (MI) classifier is:

\[
\text{Recall (MI)} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]
= 41 / (41 + 5)
= 89.13%

Then, the F-score percentage of myocardial infarction (MI) classifier is:

\[
F\text{-score}(MI) = \frac{2 \times \text{precision}(MI) \times \text{recall}(MI)}{\text{precision}(MI) + \text{recall}(MI)}
\]

\[
= (2 \times 100 \times 89.13) / (100 + 89.13)
\]

= 94.25%

A high percentage of F-score measure of myocardial infarction (MI) classifier indicates a high harmonic means between precision and recall.

And whereas the accuracy of the classifier is the percentage of test set instances that are correctly classified by the classifier (Jiawei Han, 2012; Sammut & Webb, 2011), then the accuracy of myocardial infarction (MI) classifier is:

\[
\text{Accuracy (MI)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}
\]

\[
= (41 + 69) / (41 + 5 + 0 + 69)
\]

= 95.65%

5.4. Evaluation results of the mobile application questionnaire:

In this section, we explain the mobile application evaluation results of the questionnaire that distributed to 43 respondents (9 experts and 34 patients) in order to measure users' satisfaction with the application and the benefit of the results given by the application to diabetics. In addition to their desire to use mobile applications to diagnose other chronic diseases.

The table 5.8 below shows the row numbers of respondents who gave each response, e.g. for the question 3, there are 28 respondents response with “Agree” that the mobile application easy to learn.
Table 5.8: Mobile application respondent’s answers

<table>
<thead>
<tr>
<th>Total</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Question</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>17</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>9</td>
<td>32</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td>28</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>10</td>
<td>29</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>19</td>
<td>23</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>24</td>
<td>16</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>30</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15.71</td>
<td>23.43</td>
<td>3.43</td>
<td>0.43</td>
<td>0</td>
<td>Average</td>
<td></td>
</tr>
</tbody>
</table>

To measure respondent’s acceptability for each question (Table 5.9), we put a weight for each response so that the weight of the answer "Strongly Disagree" equals 0, "Disagree" equals 25, "Neutral" equals 50, "Agree" equals 75 and the weight of response “Strongly agree” equals 100.

Table 5.9: Mobile application respondent’s acceptance ratio

<table>
<thead>
<tr>
<th>Acceptance Ratio</th>
<th>100</th>
<th>75</th>
<th>50</th>
<th>25</th>
<th>0</th>
<th>Response Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Agree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Question #</td>
</tr>
<tr>
<td>Agree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disagree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strongly Disagree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage</td>
<td>Acceptance Ratio</td>
<td>Average Acceptance Ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------------------</td>
<td>--------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83.14%</td>
<td>39.53 41.86 1.16 0.58 0.00</td>
<td>81.64%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78.49%</td>
<td>20.93 55.81 1.16 0.58 0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67.44%</td>
<td>2.33 48.84 16.28 0.00 0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77.91%</td>
<td>23.26 50.58 3.49 0.58 0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>85.47%</td>
<td>44.19 40.12 1.16 0.00 0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>87.21%</td>
<td>55.81 27.91 3.49 0.00 0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>91.86%</td>
<td>69.77 20.93 1.16 0.00 0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusions about the mobile application evaluation results:

- On the importance of using information technology to know the possible complications of diabetes (Question 1), the acceptance ratio is 83.14%. Some respondents believe that it is necessary to consult with the doctor to
determine the possible symptoms of diabetes, especially that there may be other factors not mentioned in the system may lead to the emergence of these symptoms.

- Regarding the ease of use of the application (Question 2), the acceptance ratio is 78.49%. Some respondents see that patients with difficulty in vision or total blindness are unable to use the application.

- About the easy of learn (Question 3), the acceptance ratio is 67.44%. 14 respondents response with “Neutral” resulting in a decrease of the acceptance ratio for that question, due to that 92.86% of them did not receive a university degree.

- About the benefit of using the application to know the possible complications of diabetes (Question 4), the acceptance ratio is 77.91%. 9.3% of those who gave response led to decrease the acceptance ratio of that question are doctors who have prior knowledge about possible complications of diabetes.

- Regarding the understanding of the results (Question 5) and the advice on using the application to know the possible complications of diabetes (Question 6), the acceptance ratio is 85.47% and 87.21% respectively. The acceptance ratios were high and convergent and this indicates the accuracy of the results and how it presented.

- For the desire for similar technology to diagnose other chronic diseases (Question 7), the acceptance ratio is 91.86%. The high acceptance ratio confirms the importance of pre-diagnosis for chronic diseases. In addition to increasing patients acceptance of the use of this technology in the future.

5.5. Summary

In this chapter, we presented the results of our experiments; we discussed the experimental environment and tools. In addition, we presented and discussed the results of the data mining system for each diabetic complications, and discuss the evaluation measurements for each classifier. Finally, we presented and discuss the conclusions of the evaluation results of the mobile application questionnaire.
Chapter 6

Recommendation and Future Works
In this chapter, we have four sections, section 6.1 presents research summary, section 6.2 presents the useful conclusions about the research experiments, section 6.3 presents useful recommendation which extracted from this thesis about integrating a data mining system with a mobile application to extract new knowledge from diabetes healthcare data, finally section 6.4 presents possible directions of future works.

6.1. Thesis Summary

In this research, we propose a data mining system integrated with a mobile application to extract new knowledge from diabetes healthcare data. The system aims to show the complications that a diabetic patient is likely to get based on the data he/she enters into the mobile application and the rules extracted from data mining system. The diabetes patients’ data were collected from UNRWA Clinics in Gaza Strip.

Our methodology in this thesis consisted of several steps: first, understand and prepared the data set collected from UNRWA clinics, in the second step we implemented data mining methods to build a classifier for extracting useful rules from the data set, and the third step, evaluated the data mining system, in the fourth step, a mobile application have been developed and used the extracted rules from data mining system to predict the complications for diabetes patients based on data that affect diabetes. The fifth step, we evaluated the mobile application.

The results of this study indicates that the average accuracy of the data mining system is 97.07%, which gives us accurate results in predicting potential complications of any new patient. As for the acceptance of specialists and patients to use the mobile application developed, the results of the questionnaire showed that the average acceptance rate of participants is equal to 81.64%.

6.2. Conclusions

The following conclusions can be drawn about the infection of diabetes complication. First, patient do not take the measurement of sugar periodically, in addition to non-adherence to the recommendations on diet, exercise, and daily lifestyle. The patient also may simply forget to take medication or measurements. In other cases, the patient may improve his lifestyle or refrain from eating saturated fat and sugars before going to the health clinic, leading to give incorrect results to specialists. Based on that, patient will be infected with serious complications of diabetes with the passage of time.

The results of this research enable diabetic patients from knowing the complications that they may get based on the data they enter into the mobile application and the rules extracted from data mining system that based on real data.
The presence of an application on a mobile device with the diabetic enable him to record blood glucose levels, calculate the rate of obesity and other factors that may result in possible complications of diabetes, forcing him to improve the lifestyle, exercise and stop smoking as well as refrain from eating Saturated fats and sugars.

However, some respondents who participated in the evaluation of the mobile application believes that it is necessary to consult with the doctor to determine the possible symptoms of diabetes, especially that there may be other factors not mentioned in the system may lead to the emergence of these symptoms. Others see that patients with difficulty in vision or total blindness are unable to use the application.

In addition, we noticed that some patients faced difficulties on learning on the mobile application because of their low university degrees. Regarding the benefit of the mobile application, we also noticed that some specialists did not receive the required benefit due to their previous knowledge about possible complications of diabetes.

As presented in chapter 3 (related work), some of the previous researches developed methods depends on provide some general tips or consults to diabetes patients on mobile application regardless of the status of each patient individually, or using data mining methods to extract knowledge from patients data without the benefit of them by patients. In contrast, our system benefit from the results of using the data mining methods to monitor and give potential complications for each patient based on his/her health status.

The mobile application and the data mining model are separated modules, resulting in a need to implement the extracted rules from the data mining model into the mobile application manually.

Due to the difficulty of obtaining medical sensors and wearable devices because of the high price compared with the per capita income in the Gaza Strip, our system is based on working only on mobile devices and this is the best solution because of the availability with the majority of patients.

6.3. Recommendations

Based on the findings and conclusion of the study, here are several recommendations to be considered:

1. Patients should change their lifestyle, stop smoking and eating unhealthy foods. In addition, they should exercise regularly to reduce the factors that lead to complications of diabetes and to maintain the sugar level within safe borders.
2. Cooperation between health institutions and agencies to carry out awareness programs aimed at raising citizens' awareness of the risks of diabetes to their health and the impact on their economic situation. In addition, ways to prevent the disease.

3. Collaborate among specialists to improve system accuracy by working to find other factors that may affect the incidence of diabetes complications.

4. Activating the work of information technology in health institutions, especially data mining, because it is important to extract hidden knowledge in all fields.

5. UNRWA and other health institutions are advised to take advantage of the system that has been carried out in this research.

6. We also recommend developing similar systems for other chronic diseases.

**6.4. Future Works**

Since this study had only focused on integrate data mining system with a mobile application to extract new knowledge from diabetes healthcare data collected from UNRWA clinics in Gaza Strip, it is recommended that future works be carried out on the following possible directions:

1. To extract new knowledge about diabetes from all other health institutions in Gaza Strip and West Bank.

2. To use other data mining methods to extract new knowledge about diabetes such as association rules.

3. To develop a mobile application using other operating systems such as IOS for Apple smartphones. In addition, to take into account differences in sizes of smartphones and tablets.

4. To apply extracted rules from data mining system directly to the mobile application, since we use in this research a rule extractor Java API routines to implement the results out from data mining system in the mobile application.

5. To develop a mobile application take into account patients with special needs and the elderly.

6. To develop similar systems for other chronic diseases.

7. Connect the system with wearable devices such as systems to measure diabetes and blood pressure.
The Reference


Bramer, M. Principles of Data Mining.


GmbH, R., RapidMiner Documentation, https://docs.rapidminer.com/ (Last Access 01/12/2017)


Jiawei Han, M. K. a. J. P. (2012). *Data Mining, Concepts and Techniques* (Third Edition ed.).


Appendices
Appendix A

Risk Scoring Protocol for Patients with Diabetes and/or Hypertension

<table>
<thead>
<tr>
<th>Major risk factors</th>
<th>Underlying risk factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Age</td>
<td>1. Obesity (BMI)</td>
</tr>
<tr>
<td>2. Elevated blood pressure</td>
<td>2. Physical inactivity.</td>
</tr>
<tr>
<td>4. Tobacco smoking any type of smoking.</td>
<td></td>
</tr>
<tr>
<td>5. Lipids disorders (Cholesterol values).</td>
<td></td>
</tr>
</tbody>
</table>

Measurable score of different risk

<table>
<thead>
<tr>
<th>No</th>
<th>Risk Factor</th>
<th>Risk Score</th>
<th>Date of assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>10-2010</td>
</tr>
<tr>
<td>1</td>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age ≤ 45 years</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>46 - 55 years</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>≥ 65 years</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Elevated BP</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Controlled with lifestyle modification*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Controlled with medications</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Uncontrolled &lt; 160/100</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Uncontrolled ≥ 160/100</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Diabetes Mellitus</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Controlled with lifestyle modification*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Controlled with medications</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Uncontrolled with medications</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Diabetes Mellitus with Proteinuria</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Smoking any type of smoking</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Lipids disorders (Cholesterol)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cholesterol &lt; 160 mg/dl</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Cholesterol 160 - 199 mg/dl</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cholesterol 200 - 249 mg/dl</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Cholesterol ≥ 250 mg/dl</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>Obesity (BMI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BMI ≤ 29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>BMI 30 - 34</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>BMI ≥ 35</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Physical activity **</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Active</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Sedentary life</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>Family history of CVD (at young ages ≤ 55 years) ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Total Score

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>13</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Stratification of patients with Diabetes Mellitus and /or Hypertension according to the risk score

<table>
<thead>
<tr>
<th>Risk</th>
<th>Low</th>
<th>Low</th>
<th>Low</th>
<th>High</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amputation, MI, CHF, Stroke, blindness, ESRF</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

* Patients on prophylaxis treatment, i.e. secondary prevention strategy.
** 20-30 minutes of activities (e.g. walking) 3 minutes per week.
*** Close relatives with hereditary link (parents, brothers, sisters, uncles and aunts) with cardiovascular diseases including strokes, MI and hypertension.
Appendix B

Easy Rule Java API sample code

class Blindness
{
	public enum e_RiskFactor{Low, Moderate, High};
	public enum e_Test{Low, Normal, High};
	public enum e_Gender{Male, Female};

	priate enum e_Result{True, False};

	priate e_RiskFactor v_RiskFactor;
	priate e_Test v_Test;
	priate e_Gender v_Gender;

	priate int vi_Age;

	priate e_Result v_Result;

	public void SetRiskFactor(e_RiskFactor p_RiskFactor)
	{
		this.v_RiskFactor = p_RiskFactor;
	}

	public e_RiskFactor GetRiskFactor()
	{
		return this.v_RiskFactor;
	}

	public void SetTest(e_Test p_Test)
	{
		this.v_Test = p_Test;
	}

	public e_Test GetTest()
	{
		return this.v_Test;
	}

	public void SetGender(e_Gender p_Gender)
	{
		this.v_Gender = p_Gender;
	}

	public e_Gender GetGender()
	{
		return this.v_Gender;
	}
public void SetAge(int pi_Age)
{
    this.vi_Age = pi_Age;
}

public int GetAge()
{
    return this.vi_Age;
}

public void SetResult(e_Result p_Result)
{
    this.v_Result = p_Result;
}

public e_Result GetResult()
{
    return this.v_Result;
}

@Condition
public boolean CheckInput()
{
    boolean vb_Result = false;
    if(this.v_RiskFactor.equals(e_RiskFactor.High))
        this.SetResult(e_Result.True);
    else if(((this.v_RiskFactor.equals(e_RiskFactor.Low))&&
        (this.v_Test.equals(e_Test.High))))
        this.SetResult(e_Result.False);
    else if(((this.v_RiskFactor.equals(e_RiskFactor.Low))&&
        (this.v_Test.equals(e_Test.Low))&&
        (this.v_Gender.equals(e_Gender.Female))&&
        (this.vi_Age > 2))
        this.SetResult(e_Result.False);
    else if((this.v_RiskFactor.equals(e_RiskFactor.Low))&&
        (this.v_Test.equals(e_Test.Low))&&
        (this.v_Gender.equals(e_Gender.Female))&&
        (this.vi_Age <= 2))
        this.SetResult(e_Result.True);
    else if((this.v_RiskFactor.equals(e_RiskFactor.Low))&&
        (this.v_Test.equals(e_Test.Low))&&
        (this.v_Gender.equals(e_Gender.Male))
        this.SetResult(e_Result.False);
    else if((this.v_RiskFactor.equals(e_RiskFactor.Low))&&
        (this.v_Test.equals(e_Test.Normal))
        this.SetResult(e_Result.False);
    else if((this.v_RiskFactor.equals(e_RiskFactor.Moderate))
        this.SetResult(e_Result.False);
    else if((this.v_RiskFactor.equals(e_RiskFactor.Low))&&
        (this.v_Test.equals(e_Test.Normal)))
        this.SetResult(e_Result.False);
    else if((this.v_RiskFactor.equals(e_RiskFactor.Moderate))
        this.SetResult(e_Result.False);
    if(this.v_Result.equals(e_Result.True))
        vb_Result = true;
    return(vb_Result);
}
@Action
public boolean BlindnessMayHappen()
{
    System.out.println("~~~~~~~~~~~~~~");
    System.out.println("~~~ Blindness ~~~");
    System.out.println("~~~~~~~~~~~~~~");

    return(true);
}
package ncd_rules;

import org.easyrules.api.RulesEngine;
import static org.easyrules.core.RulesEngineBuilder.aNewRulesEngine;

public class Controller {
    public static void main(String[] args) {
        Blindness c_Blindness = new Blindness();
        c_Blindness.SetRiskFactor(Blindness.e_RiskFactor.High);
        c_Blindness.SetTest(Blindness.e_Test.Normal);
        c_Blindness.SetGender(Blindness.e_Gender.Female);
        c_Blindness.SetAge(3);

        RulesEngine rulesEngine1 = aNewRulesEngine().withSilentMode(true).build();
        rulesEngine1.registerRule(c_Blindness);
        rulesEngine1.fireRules();
    }
}