CHILDHOOD OBESITY AND OVERWEIGHT PREDICTIONS USING HYBRID DATA MINING TECHNIQUES

by

MUHAMAD HARIZ BIN MUHAMAD ADNAN

Thesis submitted in fulfillment of the requirements

for the Degree of

Master of Science

FEBRUARY 2013

ACKNOWLEDGEMENT

In the name of Allah (the Almighty, The Gracious, The most Merciful)

Alhamdulillah, thanks to Allah for giving me the strength to complete this work. Without His permission, this work can never be completed. I am grateful to my main supervisor, Assoc. Prof. Dr. Nur`Aini Abdul Rashid for her endless support, for being a very helpful mentor, and for being a resourceful person who always shows me the right path to do this work.

I owe my sincere appreciation to my co-supervisor, Mrs. Wahidah Husain who has guided me from the first day I started to work on this project until the day I finish this work. She has been a very good motivator who always gives her best support and never gives up on me. The experiences and knowledge that I have gained will be truly treasured. I am grateful to Assoc. Prof. Dr. Cheah Yu-N and Prof Dr. Lim Che Peng, who were the panels for my proposal defense. Thanks for sharing their thoughts and for helping me to improve this work and thesis. I would like to thank the Ministry of Higher Education that has supported this study by providing the ERGS Grant for this work and also for the special funding package of the mini-budget plan.

Special thanks to my beloved parents, Muhamad Adnan Che Hussin and Kamariah Yusof for their trust in me, for always providing moral and financial support, and for always being proud of what I am doing. Special thanks to my brother, Muhamad Asyraf Muhamad Adnan for his continuous supports and his ideas that helped finalize this work.

Many thanks to the School of Computer Sciences and Institute of Postgraduate Studies, Universiti Sains Malaysia (USM) for various facilities and the kind assistance provided. Finally yet importantly to all who had helped me in any ways in completing this research including my colleagues and friends.

TABLE OF CONTENTS

ACKNOWLEDGEMENT i

LIST OF TABLES vi

LIST OF FIGURES vii

LIST OF ABBREVIATION……..……...x

ABSTRAK…….xii

ABSTRACT……..……..xv

CHAPTER 1 - INTRODUCTION

1.1 Background 1

1.2 Problem Statement 2

1.3 Motivations 4

1.4 Objectives 5

1.5 Scope 5

1.6 Methodology 6

1.6.1 Risk Factors Review 7

1.6.2 Data Collections 7

1.6.3 Parameters Selection 9

1.6.4 Empirical Studies of Selected Data Mining Techniques 9

1.6.5 Proposed Hybrid Techniques 10

1.6.6 Evaluation 10

1.7 Contributions 12

1.8 Thesis Outline 12

CHAPTER 2 - RELATED WORK

2.1 Introduction 14

2.2 Data Mining 14

2.3 Importance of Data Mining in Medical Systems 18

2.4 The Utilization of Data Mining for Predictions of Health Issues 21

2.5 Summary ..... .30

CHAPTER 3 - RISK FACTORS REVIEW AND PARAMETERS SELECTION ANALYSIS

3.1 Introduction 31

3.2 Risk Factors Review 31

3.2.1 Review of Children Factors 32

3.2.2 Review of Lifestyle Factors 33

3.2.3 Review of Family and Environment Factors 35

3.3 Parameters Selection 37

3.2.1 Analysis of Children Factors 37

3.2.2 Analysis of Lifestyle Factors 43

3.2.3 Analysis of Family and Environment Factors 51

3.8 Summary 56

CHAPTER 4 - AN EMPIRICAL STUDY USING DATA MINING TECHNIQUES

4.1 Introduction 57

4.2 Dimensionality Reduction and Feature Selection 57

4.3 Normal, Obesity, and Overweight Children Predictions 63

4.3.1 Classification and Regression Tree (CART) 63

4.3.2 Support Vector Machines (SVM) 64

4.3.3 Multilayer Perceptron Neural Networks (MLP) 64

4.3.4 Probabilistic Neural Network (PNN) and General Regression Neural Network (GRNN) 64

4.3.5 Radial Basis Function Network (RBFN) 65

4.3.6 Group Method of Data Handling (GMDH) Polynomial Neural Network (PNN) 65

4.3.7 Cascade Correlation Network (CNN) 66

4.3.8 Discriminant Analysis (DA) 66
4.3.9 K-means Clustering..... 67
4.3.10 Linear Regression 67
4.4 Results..... 68
4.5 Discussion..... 70
4.6 Summary..... 72

CHAPTER 5 - PROPOSED HYBRID DATA MINING TECHNIQUES

5.1 Introduction..... 73
5.2 Using Factor Analysis to Identify Variable Interdependencies 73
5.3 NB-Tree 75
5.3.1 Proposed Framework (NB-Tree)..... 76
5.3.2 Results and Discussion..... 79
5.4 CART Selection-NB Classification 80
5.4.1 Proposed Framework (CART Selection-NB Classification) 80
5.4.2 Results and Discussion……...……...…….83
5.5 Mean-Euclidean Classification 85
5.5.1 Proposed Framework (Means-Euclidean Classification)..... 87
5.5.2 Results and Discussion……..…….……92
5.6 ROC Curve Comparisons..... 93
5.7 Performance Evaluation using the Proposed Hybrid Techniques..... 97
5.8 Threats to Validity 101
5.9 Summary..... 103

CHAPTER 6 - CONCLUSION AND FUTURE WORK

6.1 Introduction..... 105
6.2 Revisiting Research Objective and Contribution..... 105
6.3 Recommendation for Future Works..... 107

REFERENCES……109

APPENDICES

Appendix 1: Questionnaire for Data Collection ……..119

LIST OF PUBLICATIONS 125

LIST OF TABLES

Page

Table 1.1: Collected Historical Information……..8

Table 2.1: Summary of the Literature Review…..……....……28

Table 3.1: Factors Identified from the Risk Factors Review….....38

Table 3.2: Parameters Related to the Malaysian Children……..55

Table 4.1: Discretized Variables…….58

Table 4.2: Results of Normal Predictions……...68

Table 4.3: Results of Obesity Predictions.…….69

Table 4.4: Results of Overweight Predictions…….70

Table 5.1: Results of Predictions using NB-Tree……79

Table 5.2: Results of CART Selection–NB Classification…….…….83

Table 5.3: Results of Naïve Bayes Classification without CART.....84

Table 5.4: Output Values from Normal Prediction for Abnormal Samples……86

Table 5.5: Output Values from Normal Prediction for Normal Samples……87

Table 5.6: Results of Mean-Euclidean Classification…….....92

Table 5.7: Simulating the Parameters used by Zhang et al. (2009)…….……..……..98

Table 5.8: Results of Obesity and Overweight Predictions Using Simulated Parameters……..98

Table 5.9: Results of Obesity and Overweight Predictions using Simulated Parameters……101

Table 5.10: Summarization of the Results of Each Test……..104

LIST OF FIGURES

Page

Figure 1.1: Research Methodology……..……..6

Figure 2.1: Data Mining Process *(Elmaghraby et al., 2006)*……..15

Figure 2.2: Sketch of Neuron in Human Brain *(Jain et al., 1996)*……..16

Figure 2.3: Neural Network Model. Source *(Jain, et al., 1996).* ……16

Figure 2.4: Structure of a Decision Tree.……17

Figure 2.5: Bayesian Classifier Structure *(*Liangxiao, *et al., 2009)*....……17

Figure 2.6: Data From a Patient Record *(Elmaghraby, et al., 2006).* ……21

Figure 3.1: Prevalence of Obese, Overweight, and Normal Children Based on Gender ……..39

Figure 3.2: Birth Weight CDC Growth Chart *(NCHS, 2000)*……40

Figure 3.3: Prevalence of Obesity, Overweight, and Normal for LBW, Normal Birth Weight, and HBW children ……40

Figure 3.4: Prevalence of Obesity, Overweight, and Normal for the Catch-up and No Catch-up Children……41

Figure 3.5: Prevalence of Obesity, Overweight, and Normal Children for Early, Very Early, and Late Adiposity Rebound Occurrences…….…...41

Figure 3.6: Prevalence of Obesity, Overweight, and Normal for Premature and Normal Birth Children…….....42

Figure 3.7: Prevalence of Obesity, Overweight, and Normal for Bottle-Fed and Mother-Fed Children……..……..43

Figure 3.8: Prevalence of Obese, Overweight, and Normal for Children Based on Their Breastfeeding Durations…..……44

Figure 3.9: Prevalence of Obese, Overweight, and Normal Children for Their Daily Time Spent Watching Television…….……...45

Figure 3.10: Prevalence of Obese, Overweight, and Normal Children Based on Eating

Junk Food……...…….46
Figure 3.11: Prevalence of Obese, Overweight, and Normal Children for eating Snacks While Watching TV……....47
Figure 3.12: Prevalence of Obese, Overweight, and Normal Children for Weekly Fried Food Consumptions……47
Figure 3.13: Prevalence of Obese, Overweight, and Normal Children for Having Warm Supper or None……48
Figure 3.14: Prevalence of Obese, Overweight, and Normal Children for Different Sleep Hours……49
Figure 3.15: Prevalence of Obese, Overweight, and Normal Children with Outdoor Soup or Sandwich Consumptions……...50
Figure 3.16: Percentage of Obese, Overweight, and Normal Children for Different Mother`s BMI…….....52
Figure 3.17: Percentage of Obese, Overweight, and Normal Children for Different Father`s BMI……...52
Figure 3.18: Percentage of Obese, Overweight, and Normal Children Based on Their Living Area……..55
Figure 4.1: A Scree Plot That Shows the Variance % Explained By Factor vs. the Number of Factors.....59
Figure 4.2: Variables Correlations to Components 1 to 8……..60
Figure 4.3: Variables Correlations to Components 9 to 16…….....61
Figure 4.4: Variables Correlations to Components 17 to 22……..62
Figure 5.1: Factors Identified by the Factor analysis…….…….....75
Figure 5.2: Architecture of the NB-Tree……..…….……..77
Figure 5.3: Example of Classification and Regression Tree (CART)…….……..….77
Figure 5.4: Architecture of CART Selection–NB Classification.……...81
Figure 5.5: Selected Variables Using CART…….…….…82
Figure 5.6: Architecture of Mean-Euclidean Classification……...…...…….…88

Figure 5.7: Clustering the Value in the Training Set……...……...89

Figure 5.8: Process Flow of Mean-Euclidean Classification…….……90

Figure 5.9: ROC Curves for Obesity Predictions…….……...94

Figure 5.10: ROC Curves for Overweight Predictions…….96

Figure 5.11: ROC Curves of Obesity Predictions……..……...99

Figure 5.12: ROC Curves of Overweight Predictions……..……...100

LIST OF ABBREVIATION

AI	–	Artificial Intelligence
ANN	–	Artificial Neural Network
BMI	–	Body Mass Index
BN	–	Bayesian Network
CART	–	Classification and Regression Tree
CDC	–	Center for Disease Control and Prevention
CDSS	–	Clinical Decision Support System
CHD	–	Coronary Heart Disease
CCN	–	Cascade Correlation Network
DA	–	Discriminant Analysis
FA	–	Factor Analysis
FN	–	False Negative
FP	–	False Positive
GMDH	–	Group Method of Handling
GRNN	–	General Regression Neural Network
HBW	–	High Birth Weight
ID3	–	Iterative Dichotomiser 3
IVF	–	In-vitro Fertilization
KDD	–	Knowledge Discovery and Data Mining
LBW	–	Low Birth Weight
LDA	–	Linear Discriminant Analysis

MLP	–	Multilayer Perceptron
NB	–	Naïve Bayes
NB-Tree	–	Hybrid Framework of Naïve Bayes and Decision Tree
NIH	–	National Institutes of Health
NN	–	Neural Network
PCA	–	Principal Component Analysis
PNN	–	Probabilistic Neural Network
RBF	–	Radial Basis Function
RBFN	–	Radial Basis Function Network
ROC	–	Receiver Operating Characteristics
SPM	–	Sijil Peperiksaan Malaysia
STPM	–	Sijil Tinggi Peperiksaan Malaysia
SVM	–	Support Vector Machine
TV	–	Television
TN	–	True Negative
TP	–	True Positive

RAMALAN OBESITI DAN LEBIHAN BERAT BADAN KANAK-KANAK MENGGUNAKAN TEKNIK PERLOMBONGAN DATA HIBRID

ABSTRAK

Obesiti kanak-kanak telah menjadi satu epidemik global yang membimbangkan. Obesiti kanak-kanak telah dibuktikan akan berterusan sehingga dewasa. Oleh itu, ramalan obesiti pada usia muda adalah penting dan berguna. Objektif kajian ini dilakukan adalah untuk mengenal pasti faktor risiko dan faktor penghalang obesiti kanak-kanak di Malaysia; untuk memilih dan mencadangkan parameter untuk ramalan; untuk mengkaji teknik perlombongan data yang sesuai digunakan bagi ramalan obesiti dan lebihan berat badan kanak-kanak; dan untuk mencadangkan teknik perlombongan data hibrid bagi meningkatkan kepekaan ramalan. Kajian ini mengandungi lima peringkat iaitu kajian faktor risiko, pengumpulan data, pemilihan parameter, kajian empirikal, pencadangan teknik-teknik hibrid dan penilaian. Faktor-faktor yang berkait dengan kanak-kanak di Malaysia ialah: jantina, pertumbuhan pantas, kelahiran pramatang, perlantunan kegemukan, tempoh penyusuan, sup, sanwic, makanan ringan atau coklat yang dimakan di luar rumah, menonton television (TV) secara berlebihan, memakan makanan ringan, memakan makanan ringan di hadapan TV, hidangan panas sebelum tidur, memakan makanan digoreng dan buah-buahan, aktiviti fizikal, tempoh tidur, bilangan adik-beradik, dan BMI ibu bapa. Tiga teknik hibrid telah dicadangkan. Teknik yang pertama ialah hibrid Naïve Bayes (NB) dan Klasifikasi dan Regresi

Pokok (CART) di dalam rangkakerja Pokok-NB; teknik yang kedua ialah menggunakan CART untuk memilih pembolehubah dan NB untuk ramalan (Pemilihan CART-Klasifikasi NB); dan teknik yang ketiga menggabungkan pengiraan nilai purata dan pengklasifian jarak Euclidean kepada teknik yang kedua (Klasifikasi Purata-Euclidean). Parameter yang telah dicadangkan untuk ramalan ialah: perlantunan kegemukan, obesiti ibu-bapa, pertumbuhan pantas, obesiti bapa, obesiti ibu, makanan digoreng, memakan makanan ringan di hadapan TV, memakan makanan ringan, aktiviti fizikal, bilangan adik-beradik, tempoh tidur, memakan buah-buahan, kelahiran pramatang, dan tempoh menonton TV. Ramalan obesiti menggunakan parameter yang dicadang telah menunjukkan keputusan berikut: CART dengan 66.7% kepekaan; NB dengan 0% kepekaan; Pokok-NB dengan 0% kepekaan; Pemilihan CART-Klasifikasi NB dengan 40% kepekaan; dan Klasifikasi Purata-Euclidean dengan 60% kepekaan. Manakala, untuk ramalan lebihan berat badan menggunakan parameter yang dicadang, keputusannya adalah seperti berikut: CART dengan 0% kepekaan; NB dengan 0% kepekaan; Pokok-NB dengan 0% kepekaan; Pemilihan CART-Klasifikasi NB dengan 75% kepekaan; dan Klasifikasi Purata-Euclidean dengan 95% kepekaan. Ramalan-ramalan juga telah dibuat menggunakan parameter simulasi iaitu: jantina, berat semasa lahir, kelahiran pramatang, pertumbuhan pantas, dan BMI ibu-bapa. Ramalan obesiti menggunakan parameter simulasi telah menunjukkan keputusan berikut: CART dengan 33.33% kepekaan; NB dengan 0% kepekaan; dan Klasifikasi Purata-Euclidean dengan 50% kepekaan. Manakala, untuk ramalan lebihan berat badan menggunakan parameter simulasi, keputusannya adalah seperti berikut: CART dengan 0% kepekaan; NB dengan 0% kepekaan; dan Klasifikasi Purata-Euclidean dengan 50% kepekaan. Keputusan-keputusan ini telah menunjukkan bahawa teknik-teknik hibrid yang

dicadangkan (CART-Klasifikasi NB dan Klasifikasi Purata-Euclidean) dan parameter yang dicadangkan telah meningkatkan prestasi CART dan NB dalam ramalan obesiti dan lebihan berat badan.

CHILDHOOD OBESITY AND OVERWEIGHT PREDICTIONS USING HYBRID DATA MINING TECHNIQUES

ABSTRACT

Childhood obesity has become a worrying global epidemic. Evidences show that childhood obesity persists into adulthood. Therefore, predicting obesity at an early age is both useful and important. The objectives of this study are to identify significant risk factors and protective factors of childhood obesity in Malaysia; to select and propose parameters for prediction; to investigate suitable data mining techniques for childhood obesity and overweight predictions; and to propose hybrid data mining techniques to increase the sensitivity of predictions. This study consisted of five stages that are risk factors reviews, data collections, parameter selection, empirical study, propose and evaluation of hybrid techniques. The factors that are related to the Malaysian children are: sex, catch-up growth, premature birth, adiposity rebound, breastfeeding duration, soup, sandwich, snack or chocolate eaten outdoors, excessive television (TV) watching, eating junk food, eat junk food in front of TV, warm meals for supper, eating fried food and fruits, physical activity, duration of sleep, number of sibling, and parents BMI. Three hybrid techniques have been proposed. The first technique is a hybrid of Naïve Bayes (NB) and Classification and Regression Tree (CART) in a NB-Tree framework; the second technique is using the CART for variable selection and NB for prediction (CART Selection-NB Classification); and the third technique integrates mean value calculation and Euclidean Distances classification into the second technique (Mean-Euclidean

Classification). The proposed parameters for prediction are: adiposity rebound, parent obesity, catch-up growth, father obesity, mother obesity, fried food, snacks in front TV, junk food, physical activity, number of sibling, sleep duration, fruits, premature birth, and duration of watching TV. Obesity predictions using the proposed parameters have shown these results: CART with 66.67% sensitivity; NB with 0% sensitivity; NB-Tree with 0% sensitivity; CART Selection-NB Classification with 40% sensitivity; and the Mean-Euclidean Classification with 60% sensitivity. Meanwhile, for overweight predictions using the proposed parameters, the results are as follows: CART with 0% sensitivity; NB with 0% sensitivity; NB-Tree with 0% sensitivity; CART Selection-NB Classification with 75% sensitivity; and Mean-Euclidean Classification with 95% sensitivity. The predictions were also made using the simulated parameters that are: sex, birth weight, premature birth, catch-up growth, and parents BMI. Obesity predictions using the simulated parameters have shown these results: the CART with 33.33% sensitivity; the NB with 0% sensitivity; and the Mean-Euclidean Classification with 50% sensitivity. Meanwhile, for overweight predictions using the simulated parameters, the results are as follows: CART with 0% sensitivity; NB with 0% sensitivity; and Mean-Euclidean Classification with 50% sensitivity. The results have shown that the proposed hybrid techniques (CART Selection-NB Classification and Mean-Euclidean Classification) and the proposed parameters have increased the performance of the CART and the NB in obesity and overweight predictions.

.

CHAPTER 1

INTRODUCTION

1.1 Background

Obesity is defined as abnormal or excessive fat accumulation that may impair health (WHO, 2011b). Body fatness of a person can be indicated by using the Body Mass Index (BMI). Based on the Center for Disease Control and Prevention (CDC) growth chart, a child with a BMI above the 95th percentile is classified as obese and BMI values of between 85^{th} to 95^{th} percentile is classified as overweight.

Childhood obesity has become a very worrying global epidemic where in the year 2010, there were around 43 million overweight children under the five years old (WHO, 2011b). Overweight and obesity that were once considered as the developed country problem is now rising in poor and developing countries, especially in urban settings (WHO, 2011b). Obesity that occurs during childhood will eventually persist into adulthood and lead to many co-morbid disorders such as type 2 diabetes, cardiovascular, early puberty, suicide, smoking, bulimia, poor self-image, social isolation and hypertension (Schneider & Brill, 2005).

According to Noor (2002), childhood obesity has become a notable phenomenon in Malaysia. Maso (2011) stated that the prevalences of childhood obesity in primary schools in four of Malaysia Peninsular Regions were 6% of both boys and girls. She also mentioned that the children aged 7 and 10 years in Selangor have indicated 6.6% and 13.8% prevalence of obesity respectively. Maso (2011)

also stated that in 1990 to 1997, the prevalence of childhood obesity in Klang Valley was increased from 1% to 6%.

Obesity can be prevented from a very young age if the causing factors were known. The study to identify the factors of childhood obesity and prevention in the medical area uses many tools and methods. One of the methods to prevent childhood obesity is by predicting the potential of a child to be obese or overweight before it happens. Prediction in other words is an advanced indication based on observations, experiences, or scientific reasons (Park, 2007). Beside from medical methods, prediction of childhood obesity was also possible using data mining.

Luo (2008) defined data mining as a process of obtaining information by analyzing and summarizing data from different perspectives. Luo (2008) also mentioned that the usages of data mining include detecting patterns and extracting hidden information from data. Luo (2008) stated five tasks of data mining which are summarization, classification, clustering, association and trend analysis.

This thesis aims to identify significant factors of childhood obesity from the Malaysian children based on global childhood obesity problem. The identified factors will be used as parameters for prediction of childhood obesity and overweight of the Malaysian children. The next aim of this thesis is to apply hybrid data mining techniques for childhood obesity prediction in order to increase the sensitivity of the prediction.

1.2 Problem Statement

Researchers worldwide have studied the factors of childhood obesity in young children. There are many journals and papers that have been published regarding

childhood obesity study that include the journals and papers from Japan, Germany, India, and the United States (Sugimori et al., 2004; Kleiser et al., 2009; Apfelbacher et al., 2008; Kuriyan et al., 2007; Reilly et al., 2005). However to our knowledge, there are only a small number of similar publications from Malaysian researchers, studies originating from Malaysia, or childhood obesity study that focused on the Malaysian children.

In order to predict childhood obesity in Malaysian children, we have to refer only from global sources. This may cause wrong prediction because of different factors of childhood obesity in different regions, cultures and ethnics. Therefore, precise information and data on childhood obesity from the Malaysian children is important and highly needed for prediction in Malaysia.

Technically, data mining techniques were able to produce satisfactory prediction results in many applications but it has failed to show consistency in health related predictions (Gregori et al., 2011; Almeida, 2002; Tafeit & Reibnegger, 1999; Tu, 1996). Zhang et al., (2009) comparison have shown that widely-used data mining techniques such as the Artificial Neural Network (ANN), Bayesian Network, Naïve Bayes, Support Vector Machine (SVM) and decision tree have not produced satisfactory accuracy in childhood obesity predictions. Zhang et al., (2009) showed that the highest sensitivity of childhood obesity prediction using the Naïve Bayes was 62%. Zhang et al. (2009) also showed that the highest sensitivity for overweight predictions using the SVM was 60%. Clearly, the accuracy of predictions was not satisfactory.

Hence, based on the above-mentioned situations, this research focuses on addressing the following questions:

1. What are the factors of childhood obesity in the Malaysian children?
2. Which data mining techniques are suitable for childhood obesity prediction?
3. How can the performance (sensitivity) of the identified prediction techniques be improved for childhood obesity prediction?

1.3 Motivations

For the past few years, the number of obesity cases was rising. It has increased the awareness and efforts by the authorities and researchers around the world to observe and monitor the children from an early age (Danielzik et al., 2004; Kleiser et al., 2009; Apfelbacher et al., 2008; Toschke et al., 2008; Delgado-Noguera et al., 2009; Rey-Lopez et al., 2008; Sugimori et al., 2004; Kuriyan et al., 2007). Childhood obesity has also become notable in Malaysia (Noor, 2002). Hence, this study is motivated to prevent childhood obesity for the Malaysian children.

The data mining techniques can be applied in childhood obesity predictions. The data mining techniques are suitable to be integrated in a medical decision support system or medical information system. Online medical information systems have existed and widely used such as the Net Doctor, National Institutes of Health (NIH), and WebMD. Hence, this study was motivated by the existing information system and our aim is to investigate suitable algorithm for future usage in the medical information systems for childhood obesity prediction.

1.4 Objectives

Based on the problems and motivations, this thesis seeks to fulfil the following objectives:

1. To identify significant risk factors and protective factors of childhood obesity related to the Malaysian children.
2. To select and propose parameters for childhood obesity prediction.
3. To investigate and identify suitable data mining techniques for childhood obesity prediction.
4. To propose a hybrid algorithm using the data mining techniques identified in 3 to increase the sensitivity and has good specificity of childhood obesity prediction.

1.5 Scope

The scope of this study is as follows:

1. This study focuses on identifying risk factors and parameters of childhood obesity predictions using the data from local children.
2. This study uses the data that were collected from children age 9 to 11 years old in Malaysia. The children were not monitored from birth until 5 years old. But from the data collection, we tried to obtain the information of the children from birth until 5 years old.
3. This study only uses a small size of data that sufficient for prediction.
4. This study investigates suitable data mining techniques for childhood obesity prediction from the literature survey and empirical studies.

1.6 Methodology

The methodology consists of five stages as shown in Figure 1.1.

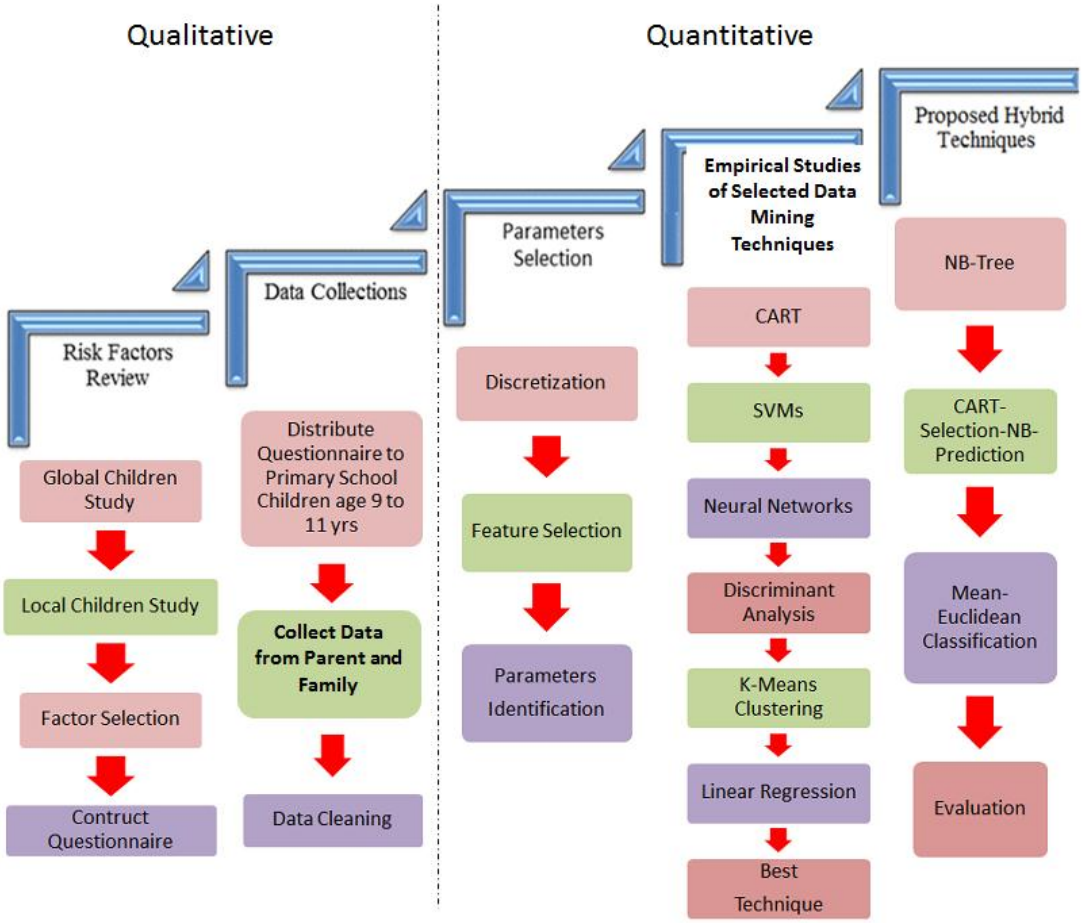


Figure 1.1: Research Methodology

The first stage of this study was the risk factor review. It involves the study of global and local childhood obesity and overweight from secondary sources such as the journals and proceedings. From the review, factors related to childhood obesity and overweight were identified. Then, questionnaires were constructed using the identified factors (Appendix 1). Each factor was transformed into questions. Then, data were collected in the secondary stage. Data collection aims to collect important information about the Malaysian children at primary school. The target group was children age 9 to 11 years old to obtain data of the children, parents, and family. The

risk factors review and the data collection stages are the qualitative part of the research.

Then, the next stage was parameters selection. It identified significant parameters from the collected data through analysis and feature selections. The fourth stage was the empirical study, where several data mining techniques were applied to childhood obesity predictions. Based on the literature review and the empirical study results, hybrid data mining techniques were proposed in the last stage. Three hybrid techniques were proposed and evaluated at this stage. The empirical study and the propose hybrid techniques are the quantitative part of this research. The details of each stage are presented in the subsequent sections.

1.6.1 Risk Factors Review

This stage identified the risk factors and the protective factors of childhood obesity from the secondary sources such as the journals, proceedings, and websites. These factors were categorized into three categories; children profile, lifestyle, and family/environment based on the study by Novak & Bigec, 1995. The review is presented in details in Chapter 3. The identified factors were transformed into questionnaire for data collections.

1.6.2 Data Collections

The questionnaires were distributed to primary school children aged 9 to 11 years old in Malaysia. Three hundred and fifty sets of questionnaires were distributed to the children. The data collection aims to collect information about the children, their

family, and environment historically from birth until 5 years old. The information details can be seen in Table 1.1.

Table 1.1: Collected Historical Information

Information	Details
Catch-up growth	From birth to 6 months old
1. Adiposity rebound 2. Duration of breastfeeding 3. Duration of sleep 4. Ate junk food 5. Duration of watching TV 6. Ate warm meals for supper 7. Ate fruit 8. Physical activities 9. Ate soup and sandwich bought outside the home 10. Parent`s BMI	From 3 years to 5 years old

The next step after the data collections processes was the data cleaning process. The data cleaning process is an important process to prepare the data for the next stage. Elmaghraby et al. (2006) and Huang et al. (2004) mentioned that the datasets can be messy with possible missing values or distortions. Han & Kamber (2001) and Kantardzie & Srivastava (2005) mentioned that the data preparation can be classified into two tasks, which are to formalize the data for processing by data mining and to prepare the data sets that maximize the performance of the data mining technique. The data were cleaned from missing and redundant values, noises, and duplications. The missing values were estimated based on the remaining values or were ignored and the attributes with many missing values were eliminated. The data cleaning process has prepared 320 datasets. From the dataset, the ratio of normal, obese, and overweight children is 6.1: 2.3: 1.6. The next stage was parameters selection.

1.6.3 Parameters Selection

This stage selects significant parameters for prediction using data analysis and feature selection. Both processes are important for this stage. The data analysis was applied using the SPSS, an IBM statistical tool for analysis purpose. The analysis is presented in detail in Chapter 3.

The feature selection was defined as a process to choose a minimal number of attributes (Pardalos et al., 2005 and Huang et. al., 2004). There is no guarantee that using more parameters will give better results than using a smaller number of parameters. The Principal Component Analysis (PCA) was used to select significant parameters. Huang et al. (2004) mentioned that improvements in the performance of most classifiers have been achieved by removing correlated or irrelevant features and in the medical research, usually only a partial attribute is used. The Factor Analysis (FA) was used to remove correlated parameters. We decided to reduce the number of parameters because the size of our data that are small and some parameters might be non-significant due to small number of data related to the parameters.

1.6.4 Empirical Studies of Selected Data Mining Techniques

The empirical study was important for this research and is presented in Chapter 4. The objective of the empirical study was to investigate suitable data mining techniques for childhood obesity prediction. Eleven data mining techniques, which grouped into six groups as follows, were applied for prediction:

- Classification and Regression Tree (CART).
- Support Vector Machines (SVM).
- Neural Networks (NN).

- Discriminant Analysis (DA).
- K-Means Clustering.
- Linear Regression.

The empirical study utilized a program known as DTREG, which was developed by Sherrod (2003). DTREG has been used in many researches to utilize the data mining techniques (Islam et al., 2010; Tooke et al., 2009; Xu et al., 2005). The best technique that was identified from the empirical study will be used to propose hybrid techniques in the next stage of the methodology.

1.6.5 Proposed Hybrid Techniques

Data mining techniques, the CART was selected from the empirical study and Naïve Bayes was selected from the literature survey. They were hybrid to propose a new algorithm for childhood obesity predictions. Other techniques that are suitable to enhance the performance of the hybrid algorithm such as the mean calculation and Euclidean distances were hybridized with the proposed technique. The proposed hybrid techniques will be discussed in Chapter 5. The evaluation criterion of the proposed hybrid techniques is presented in the next section.

1.6.6 Evaluation

The proposed hybrid techniques were utilized for normal, overweight, and obesity predictions. The results of the overweight and the obesity predictions are more important than normal prediction. The main challenge is to differentiate between

overweight and obese cases because both predictions have very similar risk factors. The evaluation parameters are specificity, sensitivity, and overall accuracy (J. Chen et al., 2007; Huang et al., 2004; Zhang et al., 2009). TP, FN, FP, and TN are defined as follows:

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

The sensitivity or the true positive rate (TPR) is TP / (TP + FN); while the specificity or the true negative rate (TNR) is TN / (TN + FP); and the accuracy is (TP + TN) / (TP + FP + TN + FN). The prediction outputs were compared to the child real BMI obtained from the data collection. In this study, the ratio of the training set and the testing set used is 60:40. The reason that we choose this ratio is because from many examples of data mining prediction techniques and tools, the 60:40 ratio is preferred (Concepts, 2004; Pujari, 2001; Ramsey et al, 2005). However in the future, we will test using other ratio such as the 10-fold Cross Validation to see the effect.

Another method used to evaluate the performance of the hybrid techniques was the Receiver Operating Characteristic (ROC) Curve (Liangxiao et al., 2009). The ROC will help to visualize which technique was best for obesity and overweight predictions. It is because the ROC curve is useful for comparing the relative performance among different classifiers (Tan et al., 2006).

Comparisons were also made with the existing parameters proposed by Zhang et al. (2009). The study by Zhang et al. (2009) was used as a benchmark

because it used many significant parameters and data mining techniques for prediction. However, the problem to compare with Zhang et al. (2009) parameter is the data that Zhang et al. (2009) used are different than ours. Therefore, similar parameters with Zhang et al. (2009) cannot be use and simulation of certain parameters to match our data has to be made. The simulation process is discussed in detail in chapter 5.

1.7 Contributions

The contributions of this thesis lies in :

1. The identified risk factors and protective factors of childhood obesity related to the Malaysian children.
2. The proposed parameters of childhood obesity prediction.
3. Identified suitable data mining techniques for childhood obesity predictions.
4. Hybrid algorithm using the data mining techniques in (3) that improves the techniques individual sensitivity and have better specificity for childhood obesity prediction.

1.8 Thesis Outline

The rest of the thesis is organized in five chapters as described below: Chapter 2 presents a review of the literature related to the study. Chapter 3 presents the research methodology that contains an explanation of each stage taken in this study. Chapter 4 presents an empirical study using data mining techniques. Chapter 5 presents the

proposed hybrid techniques and discuss the results of each technique. Finally, Chapter 6 presents the conclusions and suggestion for future work.

CHAPTER 2

Related Work

2.1 Introduction

In this chapter, the background of the data mining is presented followed by the reviews of the usage and the importance of data mining to the medical information systems.

2.2 Data Mining

Papagergiou et al. (2005) mentioned that data mining was employed to detect patterns and to extract hidden information from large databases. The data mining tasks can be descriptive or predictive (Cheng et al., 2010). Witten et al. (2011) relate data mining tasks to machine learning processes such as pre-processing, classification, regression, clustering, association rules, and visualization. Based on Luo (2008), the data mining tasks can be classified as summarization, classification, clustering, association and trend analysis. Luo (2008) defined summarization as generalization or abstraction of a set of data that results to a smaller set. Meanwhile, he stated that classification is applied to derive the class of an object based on its attributes. Luo (2008) refers clustering as identifying groups or clusters of objects with unknown classes. Fu (1997) defined association as the discovery of togetherness or the connection of objects. Meanwhile, Luo (2008) defined trend analysis as a task to find the patterns and common attributes in data that change over time.

Elmaghraby et al. (2006) stated that the data mining processes include formulating a hypothesis, collecting data, performing preprocessing, estimating the model, interpreting the model and draw the conclusions as shown in Fig. 2.1.

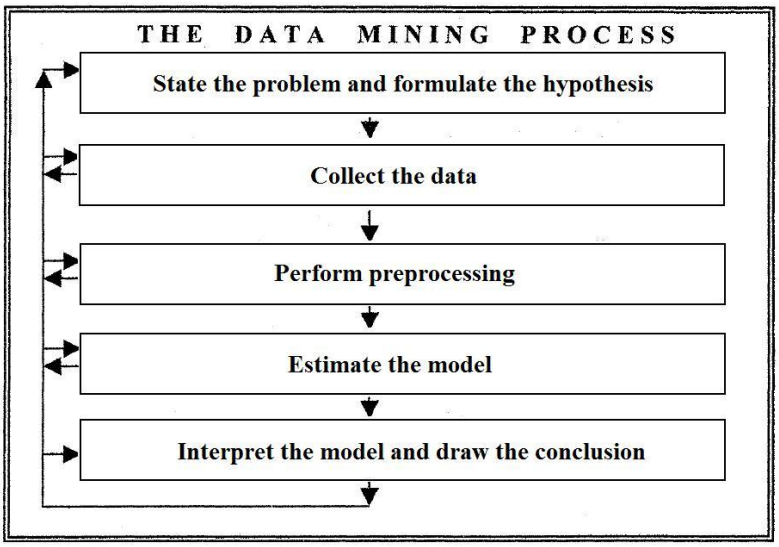


Figure 2.1: Data Mining Process *(Elmaghraby et al., 2006)*

Cheng et al. (2010) mentioned that data mining could uncover existing patterns and associations in the data or derived prediction model from the data. Many data mining techniques such as the Artificial Neural Network (ANN), decision tree, Bayesian network, Naïve Bayes (NB) and Support Vector Machine (SVM) have been applied in various domains.

The ANN is based on the biological neural networks in the human brain (Fig.2.2). Novak & Bigec (1995) described ANN as a connectionist model.

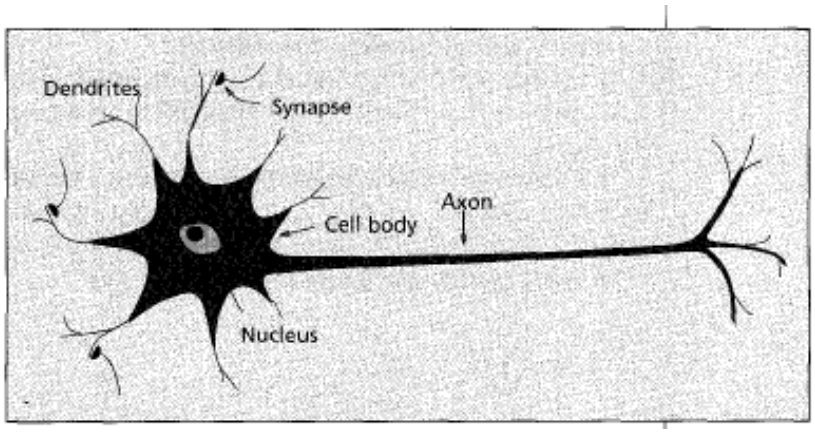


Figure 2.2: Sketch of Neuron in Human Brain *(Jain et al., 1996)*

Jain et al. (1996) mentioned that ANN was based on the neuron, a cell that processes information in the human brain. Jain et al. (1996) also explains that the neuron cell body contains the nucleus and has two types of branches: the axon and the dendrites. Jain et al. (1996) shows the connection of each neuron in Fig. 2.3.

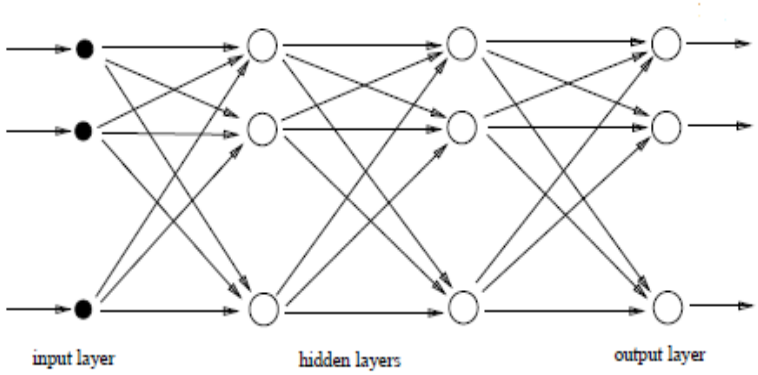


Figure 2.3: Neural Network Model (*Jain et al., 1996*)

The decision tree is a powerful classification algorithm that is popular in the information systems (Xing et al., 2007). Decision tree algorithms include the Iterative Dichotomiser 3 (ID3), assistant algorithm, C4.5, C5, and Classification and Regression Tree (CART) (Peng et al., (2009); Xing et al. (2007). Xing et al. (2007) described the decision tree was performed by separating recursive observation in

branches to construct a tree for prediction. The steps are repeated until a complete tree is built (Fig. 2.4).

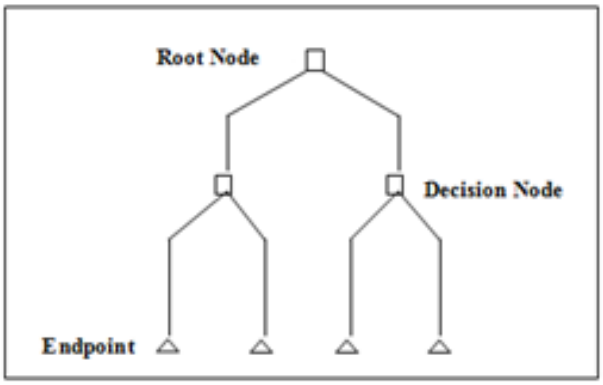


Figure 2.4: Structure of a Decision Tree

The Bayesian classifiers have a structural model and a set of conditional probabilities as shown in Fig. 2.5 (Liangxiao et al., 2009).

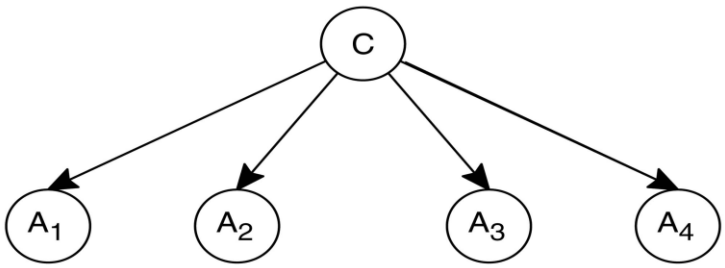


Figure 2.5: Bayesian Classifier Structure (*Liangxiao, et al., 2009*)

Zhang et al. (2009) defined the classifier of a general Bayesian network using this formula:

$$c(E) = arg\, max_{c \in C}\, P(c) P\left(a_{1,} a_{2,} \ldots, a_n \middle| c\right) \tag{2.1}$$

where a_1 *to* a_n refer to the attributes in the Bayesian network and C represents the class variable that refers to the class node in a Bayesian network. The c represents the value of C and $c(E)$ denotes the class of E. Meanwhile for Naïve Bayes (NB),

Zhang et al. (2009) mentioned that all the attributes have to be assumed as independent and therefore the formula of Naïve Bayes is as follows:

$$c(E) = arg\,max_{c \in C}\, P(c)\ \prod_{i=1}^{n} P\ (a_i|c) \tag{2.2}$$

Where the attributes a are independent attributes.

Meanwhile, J. Chen et al. (2007) described Support Vector Machine (SVM) as a classification algorithm in statistical learning theory that can provide accurate models if it can capture the nonlinearity in the data. J. Chen et al. (2007) also mentioned that the classification tasks were performed by maximizing the margin separating both classes and minimizing the classification errors. Maglogiannis et al. (2007) mentioned that the SVM training involves the optimization of a convex cost function and the learning process is not complicated by local minima. The testing used the support vectors to classify a test dataset and its performance is based on the error rate determination.

In the next section, the importance of data mining in medical systems is discussed. The reason why this study focused on data mining usage in medical systems is because childhood obesity prediction is in the area of the medical applications and the motivation of this work is to find a suitable algorithm for this kind of systems.

2.3 Importance of Data Mining in Medical Systems

Elmaghraby et al. (2006) stated that the reliance of health care on data is increasing. Therefore, medical researchers, physicians, and health care providers face the problem to use stored data efficiently when more medical information systems with large database are used (Elmaghraby et al., 2006). Lord & Wiggins (2006) stated

that the medical information system databases contained much data such as patient records, physician diagnosis, and monitoring information where the data has been useful in many medical decision support systems to save lives. Lord & Wiggins (2006) described medical decision support systems as a system that assists decision-making process and contributes to reduce medical errors and costs, earlier disease detection, and achieve preventive medicine. Examples of the medical decision support systems are Clinical Decision Support Systems (CDSS), medical imaging, and in Bioinformatics. Data mining and decision support can be integrated for better data analysis and decision support (Pur et al., 2005).

Lord & Wiggins (2006) mentioned that continuous usage of the information systems will increase the size of the database. Therefore the usage of knowledge discovery and data mining in the database (KDD) for the growing databases is important. Lord & Wiggins (2006) also mentioned that KDD gathers knowledge by identifying relations from the data sets to help predictions and its utilization is increasing in medical informatics. Prather et al. (1997); Han & Kamber (2001) mentioned that researchers had used KDD in many areas such as statistics, machine learning, intelligent databases, data visualization, pattern recognition, and high performance computing. Huang et al. (2004) stated that data mining has been used in the medical domain to improve decision making such as diagnostic and prognostic problems in oncology, liver pathology, Neuropsychology, and Gynecology.

Elmaghraby et al. (2006) mentioned that in the medical area, the task of detecting associations between risk factors and outcomes is a difficult work even for experienced biomedical researcher or health care manager. He also mentioned that data mining usage had helped clinicians to improve their health service by assisting in detecting regularities, trends, and unexpected events from the data. Meanwhile,

Cheng et al. (2010) stated that the usage of data mining tools with advanced algorithms is popular for pattern discovery of biological data which include protein interactions, sequence and gene expression data analysis, drug discovery, discovering homologous sequences or structure, construction of phylogenetic trees, gene finding, gene mapping, and sequence alignment.

Štrumbelj et al. (2010) mentioned that machine learning was not fully accepted in the medical community because medical practitioners feel that their work is more complicated using such tools. He gave an example where different models used in healthcare applications have a different explanation, especially for model-specific methods. Therefore, important things that must be considered when developing an application for medical practitioners are simplicity and the way of explaining the decisions because even the simple techniques used for medical predictions have shown reasonable results (Gregori et al., 2011). Elmaghraby et al. (2006) also mentioned that must be able to present discovered knowledge in an easy and fast manner.

Kantardzie & Srivastava (2005) classify the data that can be captured by a patient record in three groups: a structured data, semi-structured data, and unstructured data as shown in Fig. 2.6.

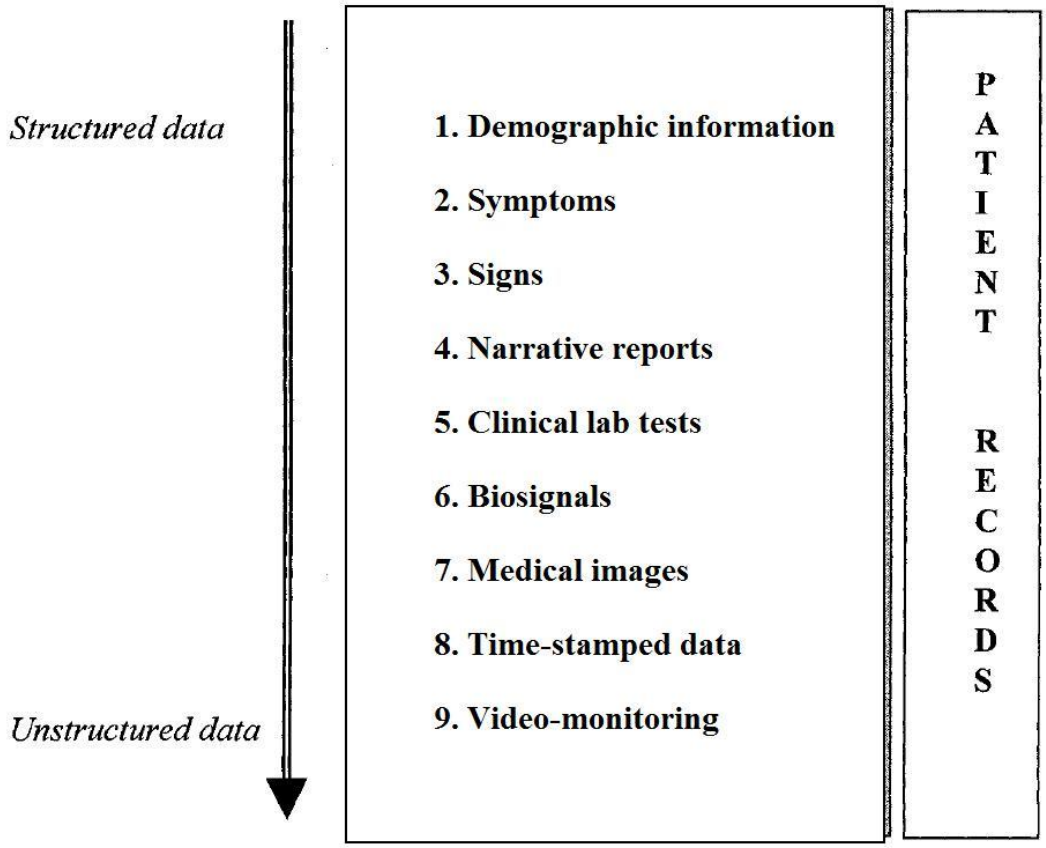


Figure 2.6: Data from a Patient Record (*Elmaghraby et al. 2006*)

Kantardzie & Srivastava (2005) also mentioned that the data mining and knowledge discovery techniques and tools based on rule induction are important to analyze the growing size of the clinical data. In the subsequent section, this thesis discusses the utilization of data mining techniques for predictions of health issues including the childhood obesity.

2.4 The Utilization of Data Mining for Predictions of Health Issues

The predictions that are discussed are coronary heart diseases, breast cancer, diabetes, and In-vitro fertilization (IVF). The similarity of these predictions with childhood obesity is the information on the negative groups (target group) is very small compared to that of the positive group. For childhood obesity prediction, the

negative group refers to the group with worse health condition that is the obese children's group and the positive group is the non-obese group. Hence, it is important to study these predictions and childhood obesity predictions to identify suitable data mining techniques for prediction.

J. Chen et al. (2007) mentioned that coronary heart disease (CHD) is a serious disease that causes many deaths especially in China. J. Chen et al (2007) also mentioned that the study on CHD patients was aimed to identify the syndrome of CHD using data mining techniques. The study used 1069 CHD cases that were collected using surveys on 5 clinical centers located in two provinces. Eighty symptoms that are closely related to CHD and always appear in the literatures of CHD were selected. Four well-known data mining techniques are tested for classifications: Bayesian network (BN), decision tree (C4.5), neural network model (MLP), and SVM. The performances of these techniques were evaluated by sensitivity, specificity, and accuracy. J. Chen et al. (2007); Huang, et al. (2004); Zhang et al. (2009) have used the sensitivity, specificity, and accuracy as follows: the sensitivity or the true positive rate (TPR) is True Positive / (True Positive + False Negative); while the specificity or the true negative rate (TNR) is True Negative / (True Negative + False Positive); and the accuracy is (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative).

The results of the CHD predictions showed that MLP accuracy was the highest (88.6%). The second best accuracy was the SVM (82.5%) and the third best was the Bayesian model (82.0%). Decision tree (C4.5) has the lowest accuracy among them (80.4%). Based on the results, the data mining technique performances were satisfactory and the accuracy of each technique did not differ much. This

proves that the data mining techniques can be used for health related prediction. The next prediction is the breast cancer prediction.

Strumbelj et al. (2010) mentioned that the study of breast cancer is interesting because of the risk factors that are difficult to be identified. Strumbelj et al. (2010) stated that the study on breast cancer recurrences involved 1035 breast cancer patients where 22 medical patient features were recorded at the time of surgery while 10 more features were recorded through follow-up. The study took more than 10 years. The data mining classifiers that were applied are the Naïve Bayes (NB), decision tree, tuned SVM, Random Forests, multilayer ANN, and bagging with NB (bag). Strumbelj et al. (2010) found that the recurrence and no-recurrence classes are almost fairly distributed. The results of the data mining technique were compared with the expert oncologist predictions. Strumbelj et al. (2010) identified that the mean accuracies of the oncologist's prediction are 0.65 and 0.63 when tested on 100 instances. Meanwhile, by using the data mining techniques, the NB and bag have shown the highest accuracy at 0.7. The decision tree and Random Forests showed slightly lower accuracy at 0.67 and 0.68 respectively.

Strumbelj et al. also tested the models on binary datasets. The bag and NB have the highest mean accuracy at 0.68 and 0.678 respectively followed by the Random Forests at 0.676, decision tree at 0.674 and ANN at 0.608. This study identified that the Naïve Bayes and bag techniques showed the best accuracy and better than the oncologist's. The results show that the data mining techniques may be more reliable than the human experts for complex problems. The next prediction is the diabetes prediction.

Huang et al. (2004) described diabetes as a metabolic disorder where the body cannot make proper use of the carbohydrate and it is greatly affected by the patient

lifestyle. The study on diabetes prediction used 2017 diabetic patient clinical information and 425 features from the database. The first step taken in the study was data pre-processing: data integration and reduction. The following step was feature selection using Relief to reduce the number of parameters. Three classification algorithms were used to analyze the data. The algorithms are the decision tree C4.5, IB1, and Naïve Bayes. The features were discretized to improve the results of C4.5. The result indicated that Naïve Bayes and IB1 accuracy have improved when the input parameters were reduced. This shows that feature selection is an important step for predictions using the data mining.

Huang et al. (2004) identified from the data that the discretized C4.5 has the best sensitivity while the Naïve Bayes has the best specificity. The Naïve Bayes has the least differences in both sensitivity and specificity. The results showed that decision tree produced high sensitivity but the Naïve Bayes has a better distribution of sensitivity and specificity. The next prediction is the in-vitro fertilization (IVF) prediction.

The in-vitro fertilization (IVF) prediction aims to identify the number of embryos to be transferred to the woman's womb and the selection of embryos with the highest reproductive viabilities (Uyar et. al, 2010). The similarity of the prediction with the childhood obesity prediction is that the positive and the negative samples are unfairly distributed. Uyar et al. (2010) mentioned that the study used 2453 records with 89% of them are positive samples. There were 18 features that are related to the embryo and the patients. Six classification algorithms were used that are the Naïve Bayes, k-Nearest Neighbor, Decision Tree, SVM, Multilayer Perceptron (MLP), and Radial Basis Function (RBF). Receiver Operating Characteristics (ROC) analysis was used to evaluate the classifier performance. The results showed that the Naïve