2020-03-24

DATA MINING BASED CRITICAL ANALYSIS AND CLASSIFICATION OF INFECTIOUS DISEASE FOR EFFECTIVE DIAGNOSIS AND TREATMENT

DESALEW, MOLALIGN

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

COMPUTING FACULTY

DATA MINING BASED CRITICAL ANALYSIS AND CLASSIFICATION OF INFECTIOUS DISEASE FOR EFFECTIVE DIAGNOSIS AND TREATMENT

BY

MOLALIGN DESALEW

BAHIR DAR, ETHIOPIA

January, 2018

DATA MINING BASED CRITICAL ANALYSIS AND CLASSIFICATION OF INFECTIOUS DISEASE FOR EFFECTIVE DIAGNOSIS AND TREATMENT

By
Molalign Desalew

A thesis submitted to the school of Research and Graduate Studies of Bahir Dar Institute of Technology, BDU in partial fulfillment of the requirements for the degree of

Master of Science in Information Technology

Advisor: Gebeyehu Belay (Dr of Eng.) Asct Prof.

Bahir Dar, Ethiopia
January 2018

DECLARATION

I declare that this thesis is my original work and that has not been presented for a degree any other university before this day and time. I have acknowledged all materials used in this work. I have not used any resources with fabrication, plagiarism or misrepresentation before properly cited or acknowledged.

Name of the student______________________________ Signature_____________

Date of submission: ________________

Place: Bahir Dar

This thesis has been submitted for examination with my approval as a university advisor.

Advisor Name: _______________________________

Advisor's Signature: ______________________________

Bahir Dar University

Bahir Dar Institute of Technology-

School of Research and Graduate Studies

Computing Faculty

THESIS APPROVAL SHEET

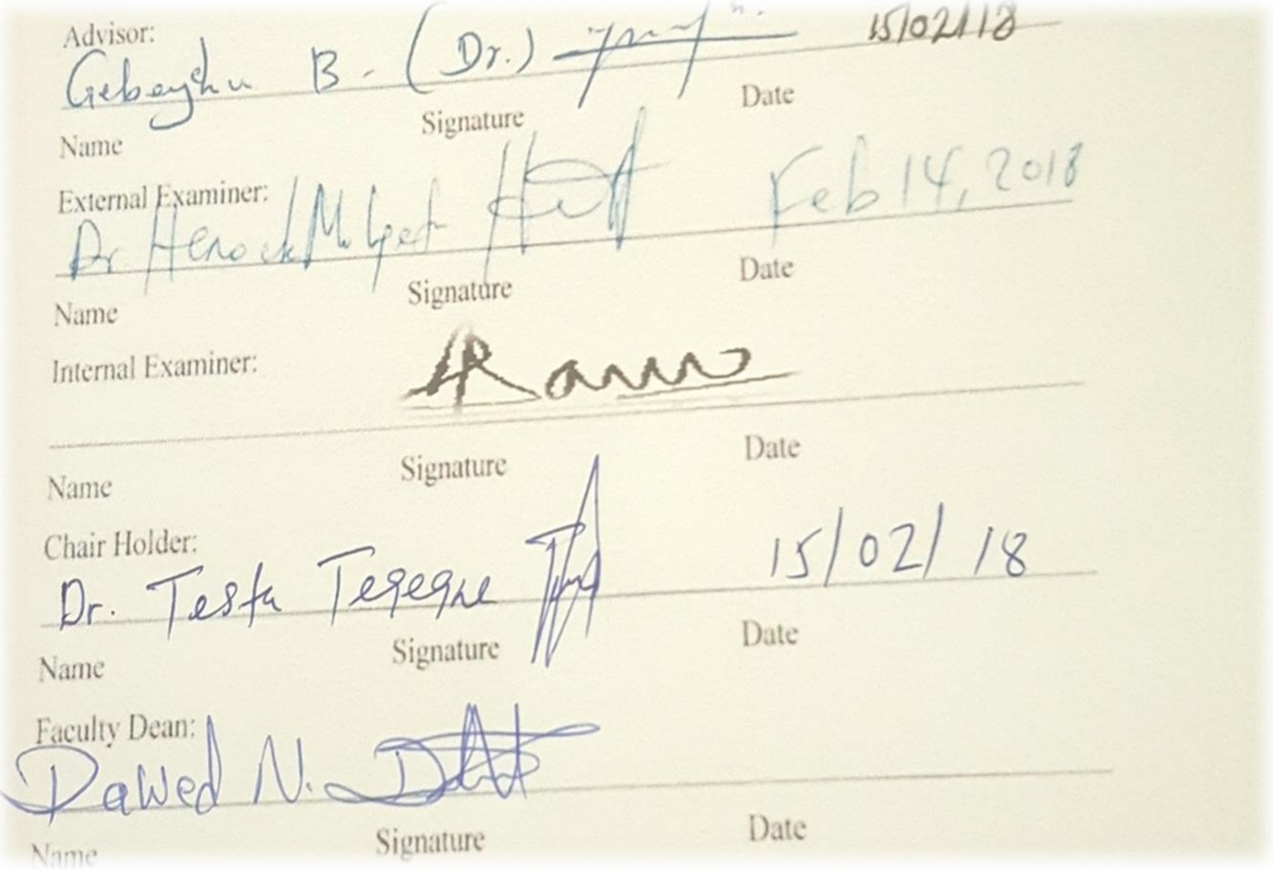
Student:

Molalign Desalew 15/02/2018

Name Signature Date

The following graduate faculty members certify that this student has successfully presented the necessary written final thesis and oral presentation in partial fulfillment of the thesis requirements for the Degree of Master of Science in Information Technology

Approved by:



Advisor:

Gebeyehu B. (Dr.) 15/02/18

Name Signature Date

External Examiner:

Dr Henock M.[illegible] Feb 14, 2018

Name Signature Date

Internal Examiner:

Name Signature Date

Chair Holder:

Dr. Tesfa Tegegne 15/02/18

Name Signature Date

Faculty Dean:

Dawed N.

Name Signature Date

Dedication

This work must be dedicated to My family, My wife and My daughters

Acknowledgement

First of all, I would like to thank my God, his wisdom and encouragement for me during this thesis work.

It is a great pleasure for me to express my heartfelt gratitude to Dr. Gebeyehu Belay, who has been an excellent advisor and mentor for me. Dr. Gebeyehu has taught me a great deal about data mining researches and guides me in a proper direction to achieve the objective of this research.

I will extend my thanks to Bahir Dar University, Bahir Dar Institute of Technology for sponsoring this thesis work. My great appreciations also give to all staffs of computing faculty for following us to complete the work on time.

I would like to give many thanks to the staffs of Felege Hiwot Comprehensive Specialized Referral Hospital and Addis Alem Hospital for allowing me to access patient medical history and assisting me whenever I needed their cooperation. I would like to thank especially Ato Tesfahun Kebede (PHEM Officer), Ato Mohammed Kassa (BSc Nurse), Dr. Niguse (AAH MD) and other staffs of the two hospitals.

Last, but not least, I am grateful for my sweet heart W/ro Mulualem Manaye whose care, tolerance, and attention made the whole process of this research work less difficult than it would be without her and my honey daughters who funs me at all times of my day and also all my families specially my mother and father their protection to reach this in my life and W/ro Shashe Takele Human Resource expert at AAH for her help at the time of the day collecting from the hospitals.

Abstract

Infectious diseases are the leading causes of illness and deaths in low and middle-income countries. They are easily transmitted from an infected individual to others through causative organisms, such as bacteria, virus, protozoa, fungi and parasites. The disease causes high morbidity, mortality and economic crisis on countries until they are treated properly. Due these cases the healthcare sectors are demanding high cost, less performance and poor patient safety care management. However, studies on such diversified issues have not done much. On the other hand, the huge amount of data is available in the health sector. Therefore, we proposed a data mining process model to extract significant patterns from patient records. We have collected a data from two government hospitals, Felege Hiwot Comprehensive Specialized Referral Hospital and Addis Alem Hospital located in Amhara National Regional State, Bahir Dar city. A total of 3017 sample datasets are collected in developing models to classify datasets of cholera, malaria and TB. The findings of this study showed that all the models built by ID3 and J48 Decision Tree, SVM and Artificial Neural Network classifiers in the three datasets has different classification accuracy. Model comparison is done based on TP (sensitivity) and FP (specificity) rates, precision, recall, F-measure, ROC area and accuracy. 10-fold cross validation test option is used to check the performances of each classifier. This suggests that the model built by the SVM classifier performs slightly better in the classification of cholera disease with a classification accuracy of 83.43%, the model built by J48 classifier performs better in the classification of malaria and TB with a classification accuracy of 98.55% and 99.2% respectively. The study showed that data mining techniques have used effectively for classifying infectious disease for effective diagnosis and treatments activities, optimize patient safety care, epidemic control and advanced preparation. The outcome of the study can be used as an inference material for physicians to support them to make more consistent diagnosis and treatments of the disease.

KEYWORDS: Infectious Disease, Data Mining, Classification, ANN, SVM, Data mining Algorithm, Knowledge

Table of Contents

Acknowledgement vii

Abstract viii

List of Abbreviations xii

List of Figures xiii

List of Tables xivv

1. Introduction 1

1.1. Statement of the Problem 3

1.2. Objectives 4

1.2.1. General Objective 4

1.2.2. Specific objectives 4

1.3. Scope of the Study 4

1.4. Contribution of the Study 5

1.5. Organization of the Thesis 5

2. Literature Review 7

2.1. Introduction 7

2.2. Data Mining in Healthcare 8

2.3. Data mining via Infectious disease datasets 11

2.3.1. Transmission methods of Infectious disease 12

2.3.2. Infectious diseases in the study 13

2.4. Data Mining Tasks in the healthcare 17

2.5. Data Mining Process Models 17

2.5.1. Knowledge Discovery in Database (KDD) 18

2.5.2. CRISP-DM Process 19

2.5.3. Hybrid Process 21

2.6. Data Mining Techniques in the Study 21

2.6.1. Classification 22

3. Methodology 23

3.1. Study Area 23

3.2. Research Design 23

3.3. Sample Size 27

3.4. Research Method…………………………………………………….........................28
3.5. Tools 29
3.6. Algorithms Used in the Conducted Research 30
3.6.1. Decision Tree Algorithm 30
3.6.2. Support Vector Machine (SVM) Classifier 34
3.6.3. Artificial Neural Network 35
4. Data Processing, Model Building and Performance Evaluation 39
4.1. Data Format 39
4.2. Data Preprocessing...42
4.2.1. Data Cleaning...43
4.2.2. Data Transformation 45
4.2.2.1. Data Discretization...45
4.2.2.2. Attribute Selection 46
4.3. Data mining and Model Development 48
4.4. Experimentation 49
4.4.1. Model Development via ID3 Decision Tree Classifier.............................50
4.4.2. Model Development via J48 Decision Tree Classifier 52
4.4.3. Model Development via Support Vector Machine Classifier...................54
4.4.4. Model Development via Artificial Neural Network classifier..................56
4.5. Empirical Analysis of Experiments 58
4.5.1. The Effect of Attribute Subset Selection 58
4.5.2. The Effect of Dataset Variation on Classifiers Performance....................61
4.6. Model Comparison..62
4.7. Generated Rules from Experiments 67
4.8. Graphical User Interface 73
5. Conclusion and Recommendation 75
5.1. Conclusion 75
5.2. Recommendation 76
References..78

List of Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
ANN	Artificial Neural Network
ARFF	Attribute Relation File Format
ASCII	American Standard Code for Information Interchange
AWD	Acute watery diarrhea
CART	Classification and Regression Tree
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma Separated Value
DACAE	Drug Administration and Control Authority of Ethiopia
DM	Data Mining
DNA	Deoxyribose Nucleic Acid
DT	Decision Tree
EDHS	Ethiopian Demographic Health Survey
EPTB	Extra-Pulmonary TB
FMOH	Federal Ministry of Health
FP	False Positive
FS	Feature Selection
HIV	Human Immunodeficiency Virus
ID	Infectious Disease
ID3	Iterative Dichotomize version 3
KBS	Knowledge Based System
KDD	Knowledge Discovery in Databases
Kg	Kilogram
MB	Mycobacterium
MLP	Multilayer Perceptron
MS	Microsoft
NR	Non-Reactive
P	Plasmodium
PTB	Pulmonary Tuberculosis
RBF	Radial Basis Function

RMSE	Root Mean Square Error
RNA	Ribbon Nucleic Acid
ROC	Receiver Operating Characteristic
R	Reactive
SEMMA	Sample, Explore, Modify, Model, and Access
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
TB	Tuberculosis
TP	True Positive
UNICEF	United Nation Children's Fund
V	Vibrio
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization
Wks	Weeks

List of Figures

Figure2. 1: Data Mining Architecture with Other Fields.. 8

Figure2. 2: Process of Data Mining in Healthcare.. 10

Figure2. 3: Infectious disease transmission flow.. 12

Figure2. 4: KDD Processes.. 19

Figure2. 5: Six Step CRISP-DM Process Models 21

Figure3. 1: Stages of Hybrid Process Model ……………………….……..………...27

Figure3. 2: General flow of the research method 29

Figure3. 3:J48 classification results of cholera dataset using tree structure.................... 33

Figure3. 4:J48 classification result for malaria dataset using tree.................................. 33

Figure 3. 5: Process description for SVM model development...................................... 35

Figure3. 6: MLP Network Structure for Malaria Dataset... 38

Figure4. 1: Data Preprocessing Tasks for Quality Data Mining…………………………43

Figure4. 2: Performance of ID3 classifier before and after attribute subset selection...... 59

Figure4. 3: Performance of MLP classifier before and after attribute subset selection.... 61

Figure4. 4: Accuracy of classifiers with 10-fold cross validation and percentage split ... 62

Figure4. 5: Classification accuracy of the four classifiers in the three datasets 66

Figure4. 6: Graphical User Interface for the Conducted Research.................................. 74

List of Tables

Table2. 1: Top 10 leading causes of outpatient visits in Ethiopia 13

Table4. 1: Cholera (AWD) Disease Attributes and descriptions…………………………...40

Table4. 2: Malaria Disease attributes and descriptions 41

Table4. 3: Tuberculosis Disease attributes with its descriptions 41

Table4. 4: Discretized attributes of Age and Physical Weight for TB patients 45

Table4. 5: Discretized attributes of Age and Body Temperature for Malaria patients 46

Table4. 6: Discretized attributes of age and diarrhea loss per day for cholera patients....46

Table4. 7: Top ranked attributes selected for model development 48

Table4. 8:ID3 classification results using all and best selected attributes 51

Table4. 9:J48 classification results by applying all and best selected attributes 53

Table4. 10: SVM classification results by applying all and best selected attributes 55

Table4. 11: MLP classification results by applying all and best filtered attributes 57

Table4. 12:ID3 classifier accuracy changes due to attribute subset selection 59

Table4. 13:J48 classifier accuracy variation on attribute subset selection 60

Table4. 14: MLP classifier accuracy variation on attribute subset selection.................... 60

Table4. 15: Performance of classifiers in 10-fold cross validation and percentage split . 62

Table4. 16: Confusion matrix for Cholera dataset from J48 classifier result 63

Table4. 17: Confusion matrix for malaria dataset from J48 classifier result.................... 64

Table4. 18: Confusion matrix for TB dataset from J48 classifier result 64

Table4. 19: Performance of all the four algorithms in the three datasets separately 65

CHAPTER ONE

1. Introduction

The dictionary meanings of diseases are a disordered or incorrectly functioning of an organ, part, structure or system of the body. It is resulting from the effect of genetics, infections, poisons, nutritional deficiency, toxicity, sickness or ailment. They attack either humans, animals or both based on the causative agents and transmission methods. It may be classified as communicable and non-communicable.

Communicable diseases are diseases that are sourced from the causative organism spreading from one person to another. They are the leading cause of illness and deaths in low and middle-income countries. These disease groups are mostly suffered sub-Saharan Africa countries due to poor sanitation and poor disease prevention practices. The diseases affect people of all ages, but highly suffer children due to their contact to environmental conditions, which support the spread. These disease groups are responsible for one quarter to one third of all deaths worldwide and cover over half of all deaths of children under the age of five. Based on WHO reports of 2012 five of the top ten causes of deaths were due to infectious disease problems and kills around 9 million peoples each year (WHO, 2012).

Communicable diseases are caused by pathogenic microorganisms, such as bacteria, protozoa, viruses, parasites or fungi. The diseases can be spread, directly or indirectly, from one person to another with various transmission methods, which include genetics, animals, food, water, air, sexual contact, body contact, sharp materials and others (Brownlie, et al., 2006).

Ethiopia experiences a heavy burden of communicable or infectious diseases. Such diseases are the leading cause of mortality, morbidity and poverty in Ethiopia. According to the Federal Ministry of Health report, communicable diseases are accounted for most of the top ten causes of illness and death in 2008/09 (FMoH, 2010). These disease problems lead the healthcare sector to generate extensive data every day. This voluminous data may be retrieved when needed, and requires critical analysis and classification to find relevant patterns for effective diagnosis, prognosis, treatment and outbreak prediction of the illness. To complete these kinds of knowledge mining tasks, data mining techniques and algorithms are applied.

Data mining is a technique, which deals with the extraction of hidden and relevant information from a huge dataset. It uses sophisticated algorithms for the process of extracting and searching of valuable information. It is an advanced and powerful technology and an interdisciplinary field of computer science, which gets high attention in the computer world (ȚĂRANU, 2015). It uses existing data as an input from different databases and transforms it into new result or information. It integrates approaches from Artificial Intelligence, machine learning and database management for the extraction of new patterns and knowledge (Gaurav & Ashwini, 2014) and (Sujatha et al, 2016).

Data mining typically generates patterns by using models, which takes input and makes one or more output through classification, clustering, acquisition to regression and prediction. The model is used for understanding the phenomena of the data, analysis of the data and prediction. To build models, data must be found as a trend data, which supports in making better decisions in the future. Now a day data mining technologies are successfully applied on various environments. For example, Bioinformatics uses it for finding of several patterns in biological data like DNA, RNA, nucleotide/protein sequence and etc (Harshank & Pushpendra, 2014).

In healthcare, data mining is becoming increasingly popular. Its applications can greatly benefit all parties involved in the healthcare industry. For example, it helps healthcare insurers detect fraud and abuse. Healthcare organizations perform various activities including improving patient treatments, disease diagnosis, outbreak prediction, applying best practices and affordable healthcare services. It is also used to filter previously unknown patterns and trends from databases and use that information to build predictive models to improve the sector (Pradhan M. , 2015).

Therefore, this research is aimed to analyze and classify infectious disease of cholera, malaria and tuberculosis for effective diagnosis and treatment activities. The outcome contributes to identify the important patient record patterns of tuberculosis, malaria and cholera datasets, and develop classify the datasets, which can be used for the diagnosis and treatments of the diseases.

1.1. Statement of the Problem

Communicable Disease has becoming an insignificant cause of illness and deaths in the developing world. Based on WHO estimation around 15 million people were died from infections, maternal health and nutritional disorders in 2010. If this number is continuing with these speeds 13 million deaths will be attributed in 2050 (WHO, 2013) and (Lozano R, et al., 2013). Developing countries are highly suffering from infectious disease problems rather than non- infectious disease cases. There is a marked difference in terms of burden of disease, morbidity and mortality among developed and developing countries. In developed countries, chronic diseases such as cardiovascular diseases, cancer and diabetes have the highest burden, and in developing countries, infectious diseases still represent the biggest issue. For example, Lower respiratory infections, HIV/AIDS, diarrheal diseases, malaria and tuberculosis (TB), collectively accounts around one third of all deaths.

As a result, the health sector generates vast amounts of datasets every day, which helps in developing predictive models and classifying the datasets for effective diagnosis, treatments, outbreak prediction, effective healthcare services, fast and improved facilities, secured information management. This data requires intelligent and effective technologies to discover hidden and unidentified information. From these technologies, data mining plays the majority paramount roles. Data mining in public health and healthcare environment is a vital field for deeper understanding of medical data.

As described above data mining is highly applied in the medical data analysis and knowledge extraction. However, most previous works are extremely concentrated on non-communicable diseases like heart disease, stroke, cancer, kidney and others. On the other hand, analysis and exploration of patterns of infectious disease are needs more emphasize. To fill this gap, a research has proposed on critical analysis and classification of infectious disease datasets using data mining approaches for effective diagnosis and treatment practices. Thus, for this research the following questions have been addressed.

- How to identify the significant parameters of infectious disease via data mining?
- Which data mining approach is suitable for the analysis and classification of selected infectious disease datasets in the study?
- What patterns are highly important for infectious disease diagnosis and treatment activities?
- Which model is shown best performance in the classification of selected infectious disease types for effective diagnosis and treatments of the disease?

1.2. Objectives

1.2.1. General Objective

- To critically analyze and classify the infectious diseases datasets of patient's, which provide effective disease diagnosis and treatment services.

1.2.2. Specific objectives

Specific objectives of this research work are.

- To identify key patterns of selected ID datasets using data mining techniques.
- To apply data mining techniques for the classification of an infectious disease for effective diagnosis and treatment services.
- To apply identified patterns through data mining techniques in new datasets of Infectious disease groups.
- To compare the models based on their classification accuracy and select the best classification model to diagnose or treat patients.

1.3. Scope of the Study

This research is aimed, to critically analyze and find important knowledge from infectious disease patient datasets for effective diagnosis and treatment by using data mining techniques. As the features and parameters of infectious diseases are heterogeneous in nature, the study is limited in classifications of cholera, malaria and TB using data mining algorithms. The key attributes for the selected disease groups are demographic information's of the patient, particular symptoms and disease complexities of each selected cases based on the severity of the disease.

1.4. Contribution of the Study

The main contribution of the study is analyzing the already existed dataset and develops a classification model of cholera, malaria and TB disease patients through data mining techniques and algorithms, which helps physicians and healthcare workers to diagnose and treat the disease effectively. The study also contributes on identifying the key parameters of the above infectious disease datasets for effective diagnosis and treatment activities in the future. This helps physicians to use the developed classification model in their day-to-day diagnosis and treatment practices of patients. Moreover, it adds quality to the decision making process for quality healthcare service. The developed classification model gives an advice for healthcare workers and other practitioners to have knowledge of cholera, malaria and TB diagnosis and treatment services using computer technologies and hospital information management experts to develop effective databases for medical and clinical datasets of patients in an organized manner for future work and decision-making purpose.

1.5. Organization of the Thesis

The study is organized into five chapters. The first chapter deals with the background of the study, which introduces data mining applications in the healthcare, statement of the problem, objective, scope and contribution of the study.

The second chapter discusses about literatures to be reviewed, which briefly discusses about data mining applications in the healthcare domain, data mining techniques to be used in the study domain and mining algorithms applied in the selected dataset samples.

The third chapter mainly focuses on the research methodology; how the research conducted, including what procedures are followed to understand the problem, collect, and analyze the data, tools to be used, and algorithms applied in the study.

The fourth chapter discuss on the detail description of the data, model developments and performance measures of the developed models, such attempted to show the task to be done to generate a better quality dataset ready to apply data mining tools and techniques. Therefore, preprocessing tasks including data cleaning, transformation and attribute selection is discussed. And also presents the experimentation done, performance evaluation and the analysis of the result using selected hybrid techniques in data mining with selected algorithms.

The last chapter focuses on making conclusions and recommendations to show further research directions in the future, which follows references that are cited by the researcher.

CHAPTER TWO

2. Literature Review

2.1. Introduction

Data mining is the computing processes of discovering patterns, hidden information and unknown data, relationships and knowledge by exploring the large datasets. It is a confluence to machine learning, statistics, Artificial Intelligence, database and others. It requires to analyze large-scale data, which cannot handle by traditional statistical methods (Jiawei al et, 2012). It is an essential process where intelligent methods are applied to extract patterns and is a key step in the overall process of knowledge discovery in databases (Sushmita & Tinku, 2003) and (Mary & Obenshain, 2004). It incorporates analytical techniques drawn from a range of disciplines such as AI, pattern recognition, statistics, visualization, machine learning and collectively examines large volumes of data and provides user friendly approaches. Data mining is the search, exploration and analysis of valuable information from a volume of datasets in order to discover patterns and rules, through the use of automatic or semiautomatic tools (Joyce, 2002).

The term patterns indicate models and regularities, which can be observed within the data. Patterns have to be valid, means they should be true for new data to some degree of certainty (Michael & Gruenwald, 1999). Data mining has grown and continues to grow, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, Artificial Intelligence, knowledge acquisition for expert systems, data visualization, and high-performance computing. The merging goal is extracting high-level knowledge of low-level data in the context of large datasets (Pradhan, 2014). It uses sophisticated algorithms for the process of selecting relevant information from huge datasets by combining methods from different disciplines such as databases, statistics, neural network, pattern recognition, information retrieval and others. It is an advanced and discovery driven process and is also essential to ask, how the data are stored and accessed, how algorithms can be scaled to massive datasets and still run efficiently, how results can be interpreted and visualized, and how man- machine interaction can usefully be modeled and supported (Gaurav & Ashwini, 2014).

Data mining provides the methodology and technology to transform massive amounts of data into useful information for decision making. The analytical objective of data mining is organizing the data, text and images of huge data into knowledge using the privilege of computers. Figure 2.1 showed the connections of data mining with other fields.

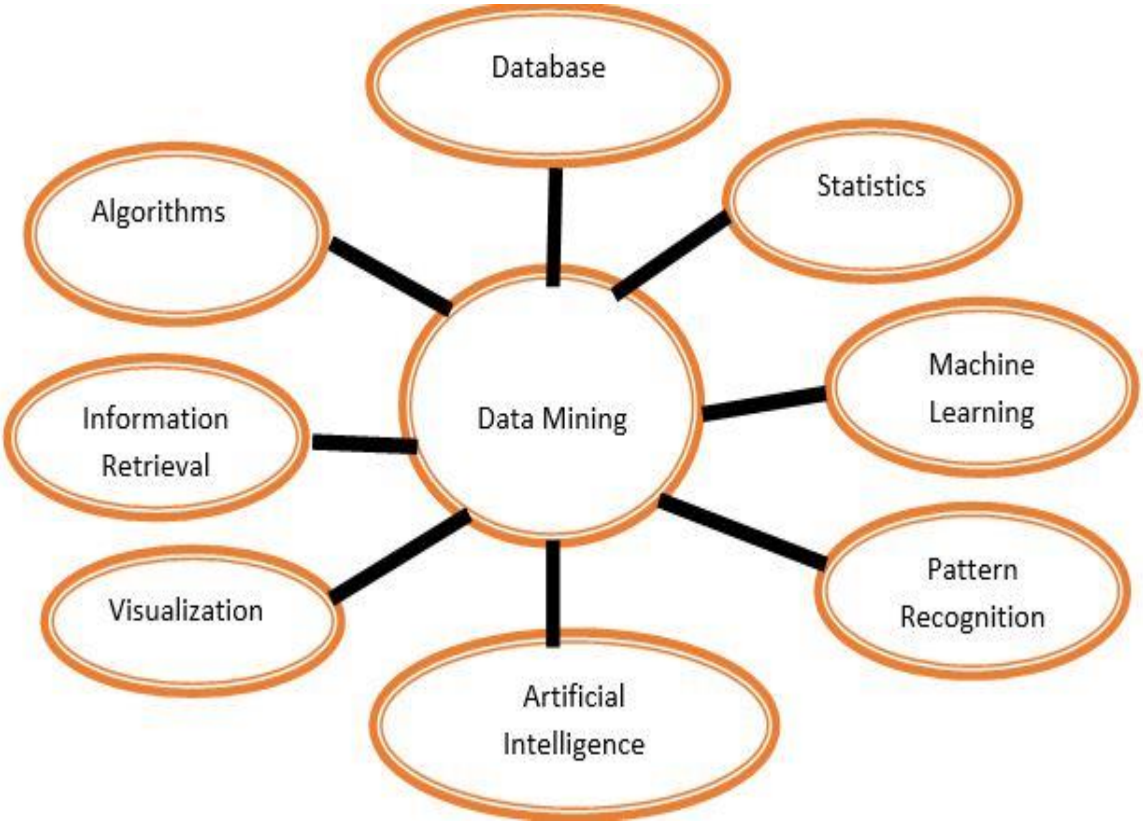


Figure2. 1: Data Mining Architecture with Other Fields

2.2. Data Mining in Healthcare

It is well known that healthcare is a complex area in which new information is being collected daily in a growing rate. An enormous part of this information is found in the form of paperwork. However, making this information available in electronic form and converting the information into knowledge is not an easy task. All healthcare institutions need an expert analysis of their medical data, which is time consuming and expensive for human analysts (Durairaj & Ranjani, 2013). The ability to use a data in databases in order to extract useful information for quality healthcare service is a key for the success of all healthcare institutions (Divya & Sonali, 2013).

Healthcare data, contains all the information regarding to patients as well as the parties involved in healthcare industries. The size and complexity of such data are increased from time to time. Due to this, the healthcare sectors need large storage places for their data as well as intelligent technologies to retrieve meaningful information from the complex and dirty dataset collections (Parvez, Saqib, & Syed, 2015). Using traditional methods for extracting meaningful information from these complex datasets is impossible. However, due to advancements in the fields of statistics, mathematics, databases and data mining disciplines it is possible to extract meaningful patterns from it. Data mining is slowly but increasingly applied to overcome various problems of knowledge discovery in the healthcare (Ruben, 2009).

Data mining is widely applied in the area of genetics and medicine, which is called medical data mining. There is an explosive growth of biomedical data, ranging from those collected in pharmacological studies and cancer therapy, kidney, heart and brain disease investigations to those identified in genomics and proteomics research. The rapid growth of biotechnology and biological data analysis methods are needs a new field to discover interesting and important patterns from this huge data for better decision making (Pradhan, 2014). Thus, the current progresses in data mining research led the development of many efficient and scalable methods for discovering interesting patterns and knowledge from large databases, ranging from efficient classification methods up to clustering, outlier analysis, frequent, sequential and structured pattern analysis methods and visualization tools (Tasha et al, 2012).

Medical databases are collected and stored large quantities of information about patients and their clinical conditions. Relationships and patterns hidden in this huge dataset collections of the sector provides new medical knowledge, which helps physicians for disease diagnosis, prognosis, treatment and outbreak prediction, clinics and hospitals to adapt new medical practices, fast and effective medical service, practitioners and policy makers to identify and develop timely guidelines and medical insurers to provide effective services for customers (Mariammal et al, 2014). Analyzing medical datasets are improved the healthcare by enhancing the performance of patient management tasks like grouping the patients having similar type of diseases or health issues, making better diagnosis and effective treatments, predicting the length of stay of patients in hospital, designing plans for effective information management system, analyzing the various factors that are responsible for diseases transmission, helping patients to identify healthcare institutions that provide services with minimum cost, and accessing latest information about different types of diseases (Mary K. , 2004). Figure 2.2 showed the overall process of data mining in the healthcare (Kamran, 2013).

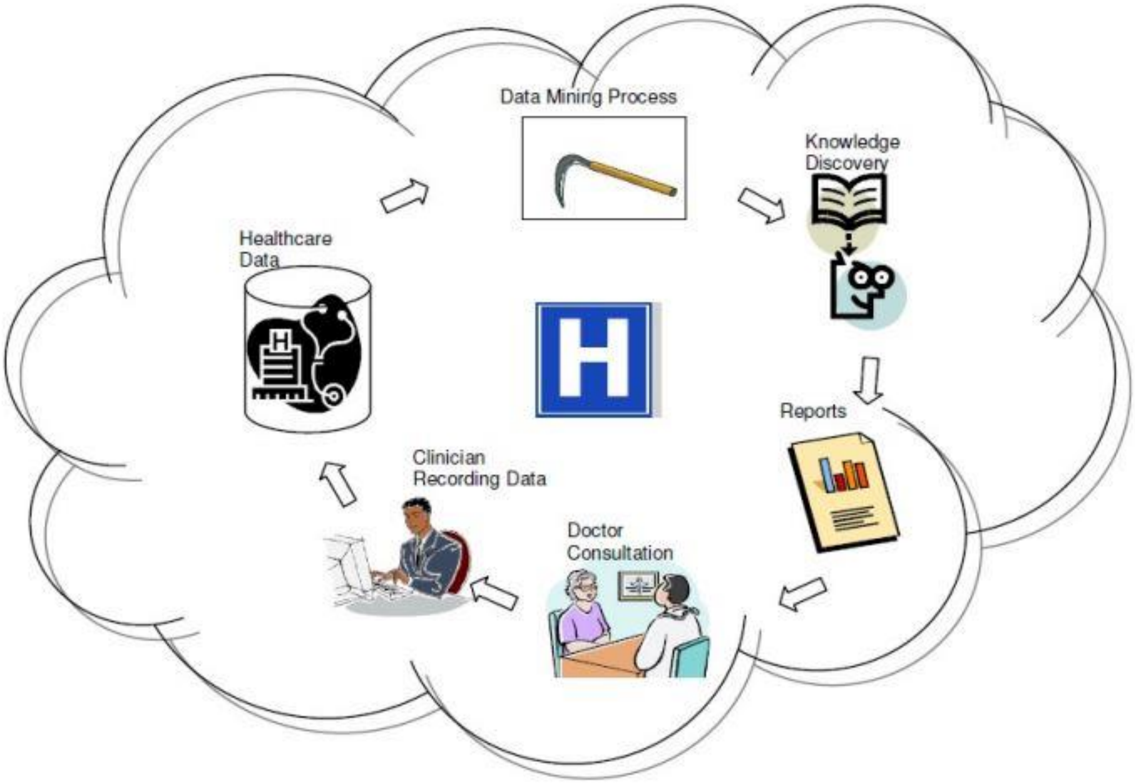


Figure2. 2: Process of Data Mining in Healthcare

The healthcare data contain details about the hospitals, patients, medical claims, treatment cost, clinical diagnosis reports, pharmaceutical prescriptions, etc(Mary K., 2004). Huge amounts of healthcare data need to be converted into information and knowledge, which can help to control costs and maintain high quality of patient care. Without data mining it is difficult to understand the full potential of data collected from a healthcare organization, which is massive, highly dimensional, distributed and uncertain. The large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost savings and decision making by discovering patterns and trends of the dataset (Desikan, Hsu, & Srivastava, 2011).

2.3. Data mining via Infectious disease datasets

Infectious disease, its name indicates a disease that is easily transmitted from person to person or animal to person through different transmission mechanisms. Infectious and parasitic diseases are major causes of morbidity and mortality worldwide. The diseases are killing more than 9 million peoples each year and many of them are children under the age of five, and poor people’s living in low and middle-income countries. They also cause enormous burdens and needs high potential to prevent them (WHO, 2012).

According to Federal Ministry of Health, Ethiopia (FMoH, 2010), communicable diseases caused by bacteria, viruses, protozoa, fungi, and parasites are highly contribute on the burden of disease, disability and death. The six leading infectious disease groups such as acute respiratory infections, HIV/AIDS, diarrheal diseases, tuberculosis, malaria and measles together causes nearly for 11 million deaths worldwide every year, and disease on the lives of tens of millions who are living with chronic or recurrent effects.

As (Mary & Sandra, 2006), studied communicable diseases have not been defeated. The microorganisms that cause these diseases are continuing to grow, and sometimes they required new drugs and methods to fight them. New pathogens arise or make the jump from infecting animals to infecting humans. The most recent global reports showed that communicable diseases are caused about one-third of all deaths.

(Alessio et al, 2015), reported that the burden of communicable diseases has been decreased gradually in the developed countries, but it is still most significant issue in low and middle-income countries. In the report lower respiratory infections, diarrheal disease and HIV/AIDS are grouped among the top major killer disease in 2011. According to (WHO, 2017), report communicable diseases are the top ten leading causes of death in the world in 2011, and they are generally responsible for extensive morbidity in all parts of the world. There is a clear difference in terms of burden of disease, morbidity and mortality among developed and developing countries. In developed countries, chronic diseases such as cardiovascular diseases, cancer and diabetes are dominant. Whereas, for developing countries, infectious diseases still represent the biggest issue.

2.3.1. Transmission methods of Infectious disease

Communicable diseases are transmitted by causative agents called infection agents or microorganisms, which live everywhere in our body and environment. They transmitted directly through physical contacts or indirectly through liquids, food, body fluids, contaminated objects, air borne breath, or vector-borne spread (Abdur , 2011). They are living with humans, animals, plants, soil, air and water. Some microorganisms are more pathogenic than others, which are more likely to cause disease. When the right circumstances are given, all microorganisms may cause infection (Linda et al, 2003). The transmission of infection agent microorganisms are always followed a repeating cycle. This cycle is called the transmission cycle of disease (GIDEON, 2010). There are different elements in this transmission cycle Figure 2.3 showed the cycle of infectious disease transmission:

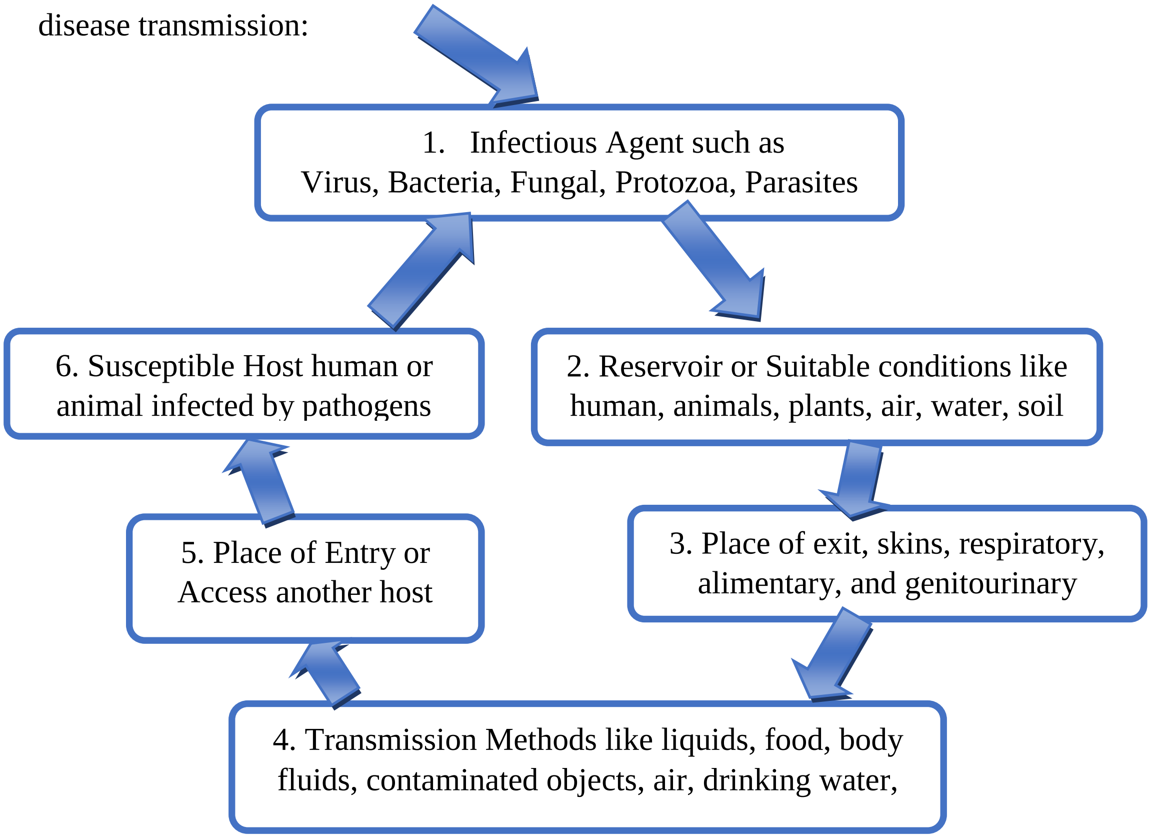

1. Infectious Agent such as
Virus, Bacteria, Fungal, Protozoa, Parasites
2. Reservoir or Suitable conditions like
human, animals, plants, air, water, soil
3. Place of exit, skins, respiratory,
alimentary, and genitourinary
4. Transmission Methods like liquids, food, body
fluids, contaminated objects, air, drinking water,
5. Place of Entry or
Access another host
6. Susceptible Host human or
animal infected by pathogens

Figure2. 3: Infectious disease transmission flow

2.3.2. Infectious diseases in the study

Communicable diseases are the major causes of health problems in Ethiopia. According to Federal Ministry of Health Ethiopian, communicable diseases are accounted for most of the top ten causes of illness and death in 2008/09. Table 2.1 shows the top 10 leading causes of outpatient visits on most regions of Ethiopia from September 2008 to August 2009 (FMoH, 2010).

Table2. 1: Top 10 leading causes of outpatient visits in Ethiopia

Rank	Diagnosis	Percentage of all outpatient visits
1	**Malaria** (clinical diagnosis without laboratory confirmation)	8.3
2	**Acute upper respiratory infections**	8.1
3	Dyspepsia (indigestion)	5.9
4	Other or unspecified infectious and parasitic Diseases	5.0
5	Pneumonia	4.8
6	Other or unspecified diseases of the respiratory system	4.0
7	Malaria (confirmed with species other than Plasmodium falciparum)	3.7
8	**Diarrhea with blood** (dysentery)	3.7
9	Helminthiasis (caused by worms)	3.5
10	Diseases of the musculoskeletal system and connective tissue (muscles, bones and joints)	3.0
Total % of all causes of outpatient visits		**47.2**

As we have shown in table 2.1, almost half of the outpatient visits on clinics and hospitals of Ethiopia are covered by transmittable infectious diseases including cholera, malaria and tuberculosis (FMoH, 2010).

Data mining is a tool used for the diagnosis of both communicable and non-communicable disease at large in the modern world. Various data mining techniques and algorithms can be applied to different diseases datasets. This helps to identify hidden associations between various symptoms on different disease types and make evidence based decisions. A number of data mining researches were carried out on various classes of biological datasets. Some of the related works carried out in mining communicable disease datasets are as follows.

(Kumar & Padmapriya, 2012), were investigated the performance of ID3 classification algorithm for the diagnosis of common disease (physical, mental, malaria, cholera and typhoid diseases) with their sign symptoms. Authors were taken a total of 600 records with 6 medical attributes from the Cleveland common Disease database. Pre-processing techniques were applied to the dataset and cluster them using the K-means clustering algorithm with K=2. The clustering result contains the data that are most relevant to common disease and the other contains the remaining data. ID3 and NN algorithms were applied to both datasets. The predicted result showed that ID3 performs higher prediction accuracy in common disease with accuracy of 94% and less prediction accuracy for non-common disease with 74% prediction accuracy.

(Mehdi , et al., 2016), were also proposed a method for detection of diseases in medical prescriptions using data mining tools and combining techniques. The study was aimed to calculate the prevalence of outpatient diseases through the characterization of outpatient prescriptions. Authors were used 1412 sample prescriptions of various types of diseases from the Iranian Ministry of Health, Food and Drug Administration. Classification algorithms like decision trees, SVM, logistic regression, NN, Naïve Bayes, Nearest Neighbor, Naïve method and Combining Techniques such as voting and stacking were used to improve the results. After classification, performance evaluation methods like accuracy, precision, sensitivity (TP rate), and specificity (FP rate) were used to test the prediction accuracy of each algorithm. Results showed that Support Vector Machine, with an accuracy of 95.32%, shows better performance than the other methods. The result of Naive method, with an accuracy of 67.71%, is 20% worse than the Nearest Neighbor method which has the lowest level of accuracy among the other classification algorithms. The result indicated that the implementation of data mining algorithms resulted in a good performance in the characterization of outpatient diseases. These results can help to choose appropriate methods for the classification of prescriptions in larger scales.

Other research is conducted by (Asia, 2012), entitled to mining patients data for tuberculosis diagnosis, in the case of Menelik II Hospital. The study was aimed to use data mining techniques for effective diagnosis of TB. The author was taken a total of 7069 sample instances from Menelik II hospital for experimentation. Two data mining techniques were applied to the data, such clustering using k-means and classification of the cluster dataset. J48 decision tree and Naive Bayes classification algorithms apply for classification of the dataset. The experiment result showed that J48 performs best classification of the dataset with a classification accuracy of 85.93%.

There was another study conducted by (Teketel, 2013), entitled constructing predictive models for occurrences of tuberculosis, the case of Menelik II hospital and ST. Peters TB specialized hospital. The objective of the study was developing a predictive model to check occurrences of TB by applying data mining techniques. The researcher used hybrid data mining process model and takes a total of 10,031 sample instances for model development. They used classification algorithms to classify the dataset. The experiment result showed that a model developed with J48 decision tree classifier has selected as a good model for predicting TB with a classification accuracy of 95.24%.

(Masumeh & Peyman, 2016), were also proposed a research work on early diagnosis of hepatitis using the hybrid model. They were using a combination of classification algorithms to get best prediction results in detection of the disease. They combine three classification algorithms, neural network, Rough set and Bayesian networks. In the result the combined approach has higher efficiency and accuracy in comparison to other methods.

(Oguntimilehin A et al, 2015), were also proposed a study which reviews a number of predictive models on the diagnosis and treatments of malaria fever. In this research work a number of conducted researches are reviewed and showed the importance of data mining techniques on the diagnosis and treatments of malaria and other infectious disease.

Another research entitled as malaria outbreak prediction model using machine learning techniques was developed by (Vijeta S et al, 2015), which was aimed to develop a predictive model for the early prognosis of malaria disease to reduce its transmission rate. In this study two popular data mining, classification algorithms Support Vector Machine (SVM) and Artificial Neural Network (ANN) were used for the development of the Malaria prediction model using a dataset of Maharashtra state of India. Root Mean Square Error (RMSE) and Receiver Operating Characteristic (ROC) were also used to measure the performance of the models. From the result the performance of the model developed using SVM is more accurate than ANN in predicting malaria outbreaks.

The other researcher (Hailu, 2015), was developed a study entitled with comparing data mining techniques in HIV testing prediction. The study was aimed to compare the prediction power of various classification algorithms to develop HIV testing prediction model. Cross-Industry Standard Process for Data Mining (CRISP-DM) model was used to design the model for HIV testing and explore association rules between HIV testing and the selected attributes. A total of 30,625 participants are involved in the study. Four popular data mining, classification algorithms Decision tree, Naive Bayes, Neural network and logistic regression were used to build the model that predicts whether an individual was being tested for HIV among adults in Ethiopia using EDHS 2011. The experimental results indicated that decision tree using the random tree algorithm performed the best with an accuracy of 96%, the decision tree induction method (J48) becomes the second best with a classification accuracy of 79%, followed by neural network (78%) and Logistic regression has the least classification accuracy of 74%.

Most of the research works are used similar data mining classification algorithms to develop the classification models, this shows the performances of the classifiers are highly different in different datasets. However, in this study different classification algorithms such as decision tree (ID3 and J48), SVM and ANN are tested for effective classification results. Different in sense the above algorithms are incomparable by its nature, support different formats of attribute values, different structure on displaying results, and have different internal default parameters.

2.4. Data Mining Tasks in the healthcare

According to (Mohammad et al, 2012) and (Sakshi & Sunil, 2015), data mining is a multistep process, which requires on accessing data and analyzing results to take appropriate action. The data to be retrieved can be stored in one or more operational database. Data mining in the healthcare has developed two types of models on the data. These are predictive and descriptive models based on the types of the healthcare data to be used for analysis and knowledge discovery purposes. In the conducted research predictive data mining approach was adopted.

Predictive (Supervised) Model: - It is also known as supervised learning. It is used to make predictions on the medical records of patients and other large scare datasets in the current voluminous data world. For example, to make a diagnosis of a particular disease, a patient must be subjected to particular treatment based on the results of treatment done on other patients with similar symptoms. To make predictions the variables under observation is split into explanatory variables and dependent variables and then, determine a relationship between the two variables. It includes tasks like Classification, regression, prediction, estimation and time series analysis.

2.5. Data Mining Process Models

There are different classes of data mining process models. Each model has its own strength and weakness. Basically, there are four types of data mining process models, which are continuously applied by researchers to discovery novel patterns and information from huge datasets. Those are knowledge discovery in databases (KDD), cross industry standard for data mining (CRISP-DM), SEMMA (Sample, Explore, Modify, Model and Access), and hybrid process models. For the conducted research hybrid data mining process model is applied, which combines both KDD and CRISP-DM features. The selected process model has high relevance for the selected research domain.

2.5.1. Knowledge Discovery in Database (KDD)

(Fayyad et al., 1996), KDD processes are multidisciplinary activities, which include techniques outside the scope of any one particular discipline. It gives special attention on finding understandable patterns, which is interpreted as useful or interesting knowledge and also emphasizes on scaling up and robustness properties of modeling algorithms for large noisy datasets. Its processes are interactive and iterative with many steps and several decisions made by the user.

According to (Abirami et al, 2013), (Pradhan, 2016) and (Shelly et al, 2011), data mining is one important method to change low level data in to high level knowledge. This is the area, which deals with the application of intelligent algorithms to get useful patterns from the data. Knowledge discovery processes are containing list of iterative and sequential steps. The processes are iterative at each step; means moving back to previous steps may be required. The process has many creative aspects in the sense that one cannot present one formula or make a complete classification for the right choices for each step and application type. Thus, it is required to understand the process and the different needs and possibilities in each step (Fayyad et al., 1996).

KDD processes are including steps such as develop an understanding of the application domain, which is the initial step in the data mining process. It refers data mining task starts by clearly understanding of the application domain with the relevant prior knowledge and identifying the goal of the knowledge discovery process. Then it follows on selecting and creating a target dataset in which, the data to be used for the knowledge discovery must be determined. This includes finding what data is available for the intended work, obtaining additional datasets if it is necessary, and then integrating all the data for the knowledge discovery activity into one dataset, by including all the necessary attributes that have important for the process. Next data cleaning and preprocessing is done on the data such as handling missing values and removal of noise or outliers. Then data transformation is applied in the preprocessed sample datasets. In this stage, the processed and cleaned data have prepared and developed for mining purpose. It includes dimension reduction (feature selection and extraction), and attribute transformation (discretization and functional transformation). This step is critical for the success of the entire KDD process, and it is usually domain specific. Next data mining tasks are completed, in which preprocessed and transformed data must be changed into meaningful knowledge by applying appropriate data mining techniques such as classification, regression, or clustering and data mining algorithms like ANN, decision tree, SVM, Naive Bayes, Fuzzy sets etc . Then, it also requires understanding the conditions under which a data mining algorithm is most appropriate for the process. Finally evaluate the discovered knowledge based on the mined patterns such as rules, methods, consistency, correctness and others with respect to the application domain and then apply the discovered knowledge to incorporate with another system for further action. The discovered knowledge is used in the existing system and measures the effects. Figure 2.4 shows the overall process of KDD from large dataset adopted from (Fayyad et al., 1996).

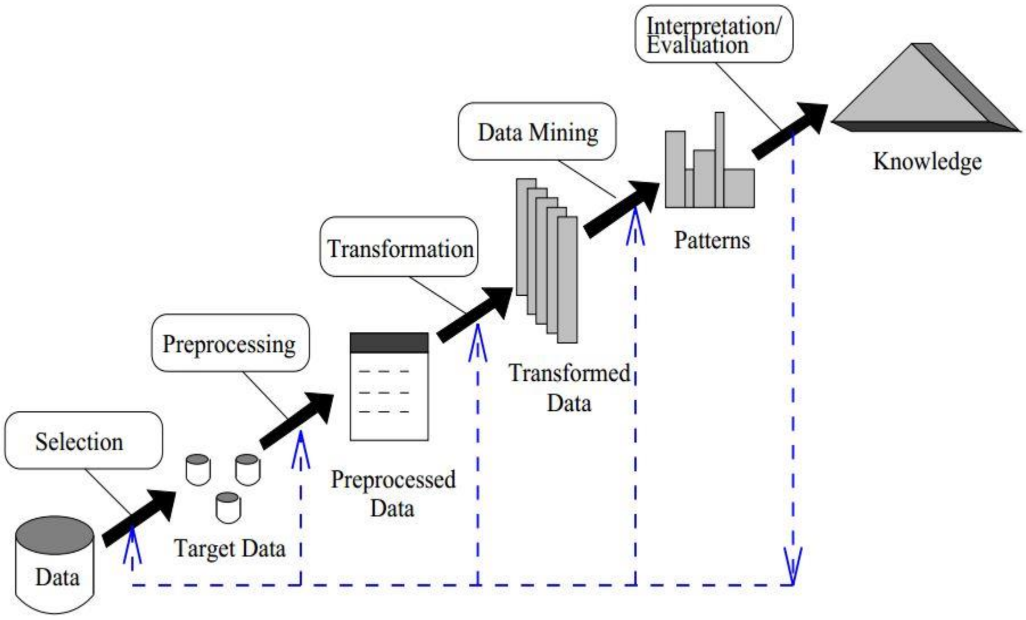


Figure2. 4: KDD Processes

2.5.2. CRISP-DM Process

Based on (Colin , 2000), (Teketel , 2013), and (Umair & Haseeb , 2014), CRISP-DM offered a general model for data or text mining projects and highlighting the key tasks of the project. It organizes the data mining process into six phases such as business understanding, data understanding, data preparation, modeling, evaluation and deployment. These phases are helps organizations to understand the data mining processes and offer a road map to follow in planning and carrying out a data mining projects. However, the sequences of the phases are not strictly applied. Moving back and forth between different phases is always required. The process is iterative because the choice of subsequent phases often depends on the outcome of preceding phases.

A CRISP-DMs life cycle, which is shown in figure 2.5 are beginning with business understanding, which focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. Then the cycle continuous with data understanding, which focuses on identifying potential input data, familiarity with the data, describe and explore the data. In this phase, high value is given for knowing the quality of data and develops initial understandings of the data. The data preparation phase includes the process of extracting relevant data for a particular modeling effort, data quality assurance, and any transformations required for specific modeling techniques. Typically, the data preparation tasks account for the majority of effort in a data mining project for data modeling. The modeling step includes selection of the modeling techniques, construction of models, generation of test designs, and assessment of models. Then evaluation of the model is conducted for all modeling techniques to check whether the developed models are achieved the business objectives or not. Finally, useful models can be embedded into information systems to support decision-making activities. The key steps are plan deployment, plan monitoring and maintenance, production of the final report, and review of the project. Figure2.5 showed the overall sequences of CRISP-DM process adopted from (Colin , 2000).

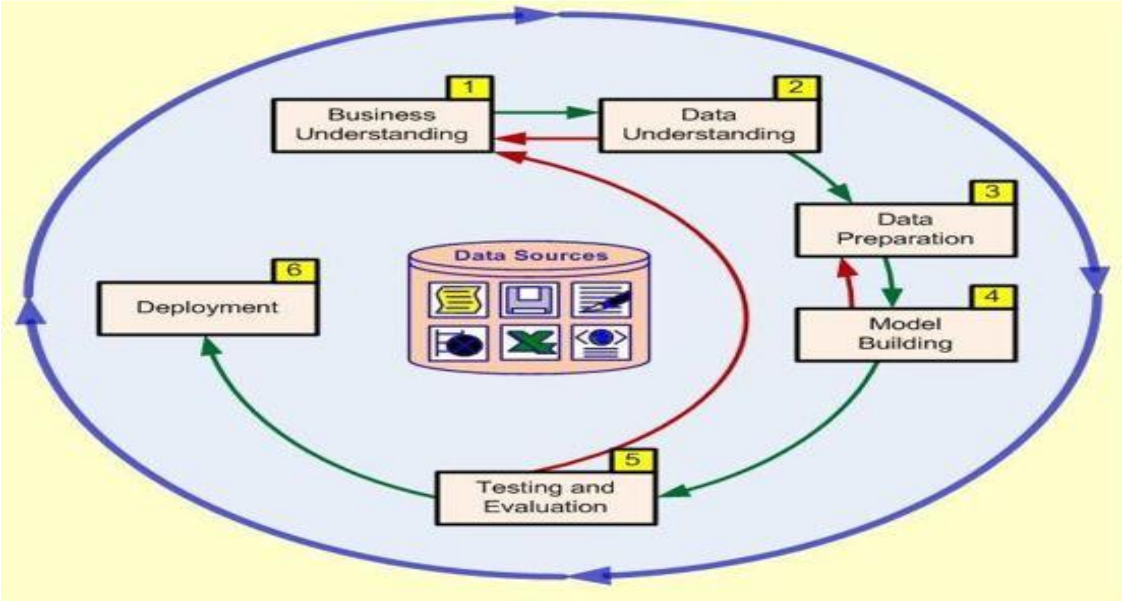


Figure2. 5: Six Step CRISP-DM Process Models

2.5.3. Hybrid Process

It is developed by combining two process models KDD and CRISP-DM. In this process model six basic steps are passed to achieve the overall goals of data mining process. It includes understanding the problem domain, understanding the data, preparing the data, mining the data, evaluating(discriminating) the discovered knowledge and finally deploy or use the discovered knowledge for real applications in the domain area. This topic is later discussed in detail in the methodology chapter.

2.6. Data Mining Techniques in the Study

The data mining techniques adopted for modeling purposes are different types depending on the types of data to be used for the development of the model. If the data mining task is predictive modeling, it permits the value of one variable to be predicted from the known values of other variables. An example of this model includes classification, regression, prediction, estimation etc. However, the mining task is descriptive modeling; it identifies patterns or relationships from data. It simply summarizes data in convenient ways or that lead to increase understanding of the way things to be work. The most common descriptive data mining techniques are clustering, summarizations, association rule mining and sequence analysis (Abel , 2011) and (Sakshi & Sunil , 2015).

In the study, the researcher is adopted classification data mining technique with three different algorithms due to the nature of the data and the objective of the study. Classification techniques are the most important approaches for the development of predictive models on pre-classified instances. In the next sub section classification data mining approach is discussed which is applied in the data in the conducted research.

2.6.1. Classification

Classification is the most commonly applied data mining technique or supervised predictive data mining technique, which contains a set of pre-classified instances to develop a model that can classify the population of records at large. The major objectives of classification techniques are developing an accurate predictive model on the pre-classified target datasets (Jiawei & Micheline, 2006).

In this research work, the researcher has adopted three classification algorithms named as Decision Tree by applying ID3 and J48 classifier, Support Vector Machine (SVM) using SMO classifier and Artificial Neural Network (ANN) using MLP classifier to develop the intended classification model, evaluate the model and compare the prediction ability of each algorithm in the three different datasets. Then categorize algorithms based on their classification accurate in the three datasets. At the end the researcher, concludes the study based on the accuracy of the developed models on the selected algorithms in the particular dataset.

The classification process includes learning the data; develop models and classifying the data by using separate training and test datasets. In the learning stage, the training data are analyzed by the classification algorithms, and in classification stage, test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data records. The classifiers training algorithms are used the existing pre-classified examples to determine the set of parameters required for proper discrimination (Asia , 2012). The algorithm then encodes these parameters into a model called a classifier. In this study, the researcher has used binary classification, which classifies cholera, malaria and tuberculosis patient datasets. This has only two possible values such as Negative refers no infectious agent and Positive refers have an infectious agent in the body.

CHAPTER THREE

3. Methodology

Research methodology is the way in which research problems are solved systematically. It is a science of studying how research is conducted scientifically. The main goals of this study is analyzing the data and develop a classification model for cholera, malaria and TB patient datasets using data mining classification algorithms. The study is conducted using hybrid data mining process model with an experimental approach, which is the combination of Knowledge Discovery in Databases (KDD) and Cross Industry Standard Process for Data Mining (CRISP-DM).

3.1. Study Area

In this study, the researcher is mainly focused on the analysis and development of classification models on infectious diseases datasets, which includes cholera, malaria and TB by taking sample data from two government hospitals Felege Hiwot Comprehensive Specialized Referral Hospital and Addis Alem Hospital, which are located in Amhara National Regional State, Bahir Dar city Administration. From these two hospitals, sample data is collected by communicating with the domain experts and information management experts, who works in the above mentioned hospitals.

3.2. Research Design

The study is designed to analyze and develop a classification model of cholera, malaria and tuberculosis patients based on the sample data, which is collected from the above mentioned government hospitals. For the conducted research, the researcher is adopted hybrid data mining process model approach to develop the model. In this study, hybrid approach is built by combining knowledge discovery in databases (KDD) and cross industry standard process for data mining (CRISP-DM) process models. The researcher selected hybrid data mining process model in conducting this research for the following reasons.

- ✓ It provides more general, research-oriented description of the steps,

- ✓ It emphasizes on the consistent aspects of the process, drawing experiences from the previous models, and
- ✓ It supports both academia and industrial data mining project standards.

In the selected data mining approach six main activities are important for the development of the required model. These include, understanding the problem domain, understanding the data, preparing the data for mining, mining the prepared data, evaluating or testing the discovered knowledge on new datasets and using or applying the discovered knowledge for decision making.

i. **Understanding the problem domain**:

In this step, the researcher is communicating and works closely with domain experts and review different documents, books, journal articles and conference papers that focus on data mining techniques and applications in the healthcare. The researcher also observes some activities of physicians in identifying the disease. This is used as a supporting source to define or formulate the problem and its corresponding solutions, determine the research objectives, identify key participants of the study, clearly understand the problem area and to select the data mining techniques, algorithms and tools suitable for solving the formulated problems.

ii. Understanding the data

Understanding the data itself is a prerequisite for any data mining technique because, the result of a knowledge discovery process highly depends on the quality of data. To understand patient medical records, a discussion is made with domain experts and identifies various attributes of the original data and then data analysis and visualization are done on the data. The major aims of this phase include, understanding the data sources, data parameters and quality of data, and identifying data format and size. The original sample data for this research were collected and described briefly. This description includes listing attributes; identifying missing values and outliers, checking completeness, redundancy and evaluating the importance of attributes to the research goal. Reduction of attributes less relevant for developing the model through attribute subset selection and the result reduces its dimensionality, and then data discretization was done for numeric attribute values to make the data nominal for all attributes. Finally, data verification is done based on the usefulness of the data in respect to the mining goal.

iii. Preparing the Data

This is the crucial phase in which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this stage, all necessary activities important for data mining are completed. It includes identification of data mining techniques and algorithms, preprocessing the sample data for mining activities, and selecting appropriate data mining tools. Data reprocessing includes data cleaning such as checking completeness of records, removing or correcting for noisy or outliers, handling missing values etc. The cleaned data were further processed by feature selection and extraction algorithms to reduce its dimensionality, and drive new attributes by discretization. Due to the nature of the datasets classification data mining technique was selected to classify the sample dataset using algorithms such as decision tree, SVM and ANN with data mining tools of Weka and rapid miner.

iv. Data Description

It is an essential research process to know the data contents more and its analysis purposes. Data mining techniques and algorithms are applied to discover interesting patterns and develop the models. This step involves usage of the planned data mining techniques, tools and selection of the new ones. Data mining tools include many types of algorithms, preprocessing techniques, and data mining elements. In the study, a classification technique is used to develop the model capable of answering the stated problems. The training and testing procedures are designed and the data model is constructed using the chosen data mining tools. The researcher used decision tree (ID3 and J48), SVM (SMO) and ANN (Multilayer Perceptron) classification algorithms, data preprocessing and model development tools such as, Notepad, WEKA 3.8 and Rapid Miner 7.4 and also documentation tools such as MS-Office packages.

v. Evaluate the Finding (Discovered Knowledge)

This step includes understanding the results, checking whether the discovered knowledge is novel and interesting, impact of the discovered knowledge and interpretation of the results by domain experts. Only the approved models are retained. The performance of each model developed in the study experiments are measured using accuracy, TP and FP rate, precision, recall, F-measure and area under the ROC. Ten fold (10-fold) cross validation performance evaluation technique was used to check the performance of the classifier in classifying the dataset and it's also compared using percentage split performance evaluation to identify the best evaluation technique.

vi. Use the Discovered knowledge

This is the final step, which consists of planning where and how to use the discovered knowledge. It determines the success of the entire knowledge discovery process. To simplify the usability of the discovered knowledge, the researcher integrated a graphical user interface to support the classification and prediction of cholera, malaria and TB status of patients. The results of the findings in this study would be reported, distributed to and used by the concerned healthcare stakeholders and other interested practitioners. Therefore, interested domain experts and researchers can get access to the research results so as to support the decision-making process, or use it for further research in the area or for any other applicable reasons. The result also used by programmers to develop a full fleshed software application for patients and physicians to easily identify the types of disease an individual has been attacked based the sign symptoms of the disease without taking any kinds of laboratory or other experimental activities (Cios et al., 2007). Figure3.1 shows the general flow of the research process and the model.

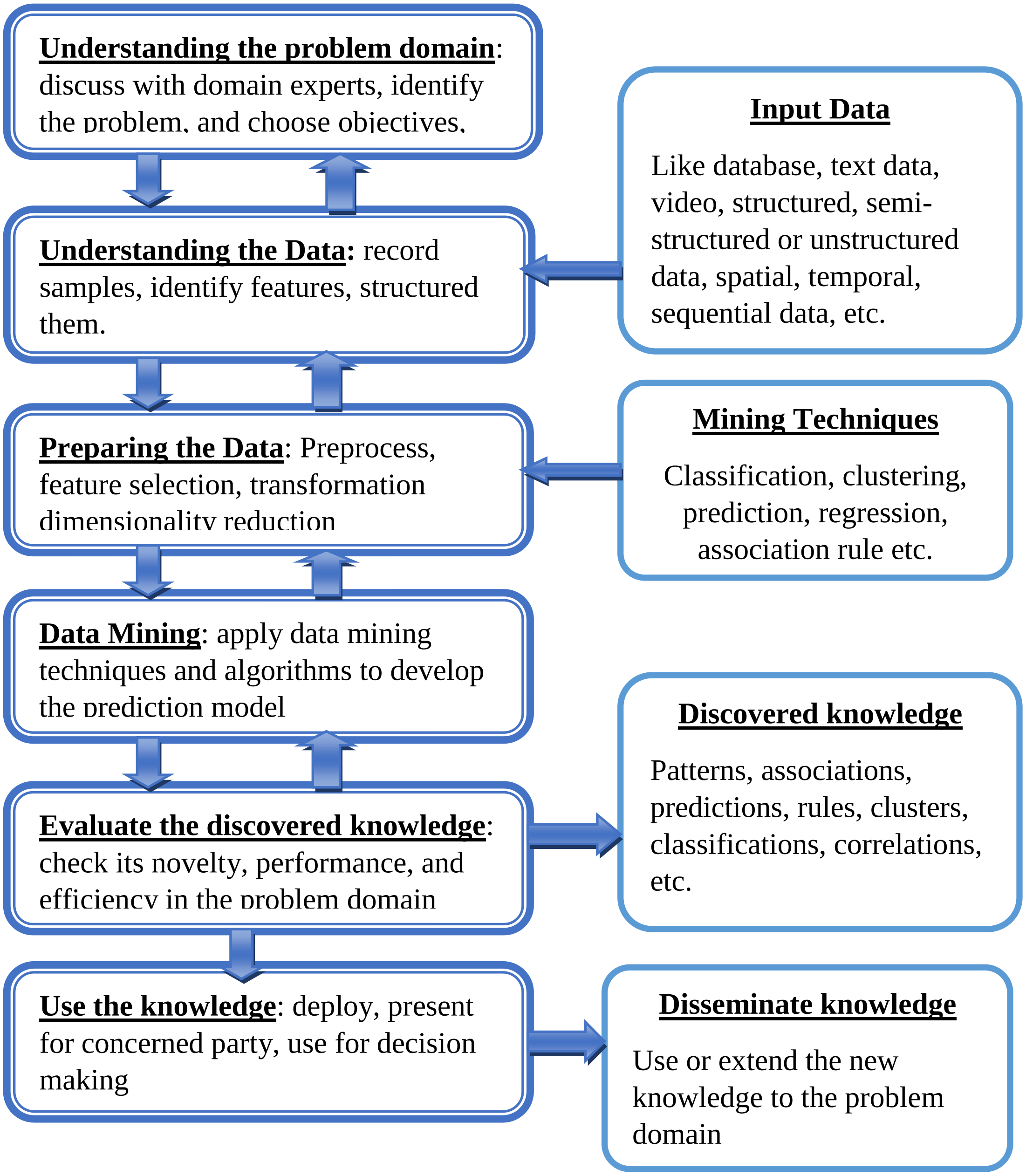

Understanding the problem domain: discuss with domain experts, identify the problem, and choose objectives,
Understanding the Data: record samples, identify features, structured them.
Preparing the Data: Preprocess, feature selection, transformation dimensionality reduction
Data Mining: apply data mining techniques and algorithms to develop the prediction model
Evaluate the discovered knowledge: check its novelty, performance, and efficiency in the problem domain
Use the knowledge: deploy, present for concerned party, use for decision making
Input Data
Like database, text data, video, structured, semi-structured or unstructured data, spatial, temporal, sequential data, etc.
Mining Techniques
Classification, clustering, prediction, regression, association rule etc.
Discovered knowledge
Patterns, associations, predictions, rules, clusters, classifications, correlations, etc.
Disseminate knowledge
Use or extend the new knowledge to the problem domain

Figure3. 1: Stages of Hybrid Process Model (adopted Cios et al., 2007)

3.3. Sample Size

The data is taken as samples from cholera, malaria and TB case datasets. A total of 3017 sample instances are taken from two government hospitals (mentioned section 3.1. above) for the development of the predictive model with 525 instances for cholera, 1100 for malaria and 1392 instances for TB cases. Each case are different in attribute types and numbers such as 10 for cholera, 16 for malaria and 15 attributes for TB cases.

3.4. Research Method

The study was conducted using a qualitative method with an experimental approach. Random sampling techniques were applied to collect sample datasets of the study. After data sampling data understanding and preprocessing is done to make the data quality for mining purposes. Preprocessing the data includes, cleaning the data such as outlier detection, removal of noisy, filling of missing values and smoothing inconsistencies, transformation such as data discretization, dimensional reduction and feature selection was done in the sample dataset collections. Then three data mining algorithms such as decision tree, SVM and ANN were applied in the preprocessed sample datasets. Next performance evaluation was done in algorithms using parameters like classification accuracy, TP rate (Sensitivity), FP rate (Specificity), precision, recall, ROC area and F-measure. Based on the performance evaluation result rules are generated based on the algorithm which performs best in the classification of the dataset(s) and then check whether the generated rules are matching with new datasets and when the rules are useful for the classification of new datasets the algorithm must be recommended for the classification of the selected infectious disease dataset for effective diagnosis and treatment. Figure 3.2 shows the general flow of the research method applied in the conducted research and selected sample datasets of infectious disease groups for effective disease diagnosis and treatment activities adopted from (Zhang, et al., 2017).

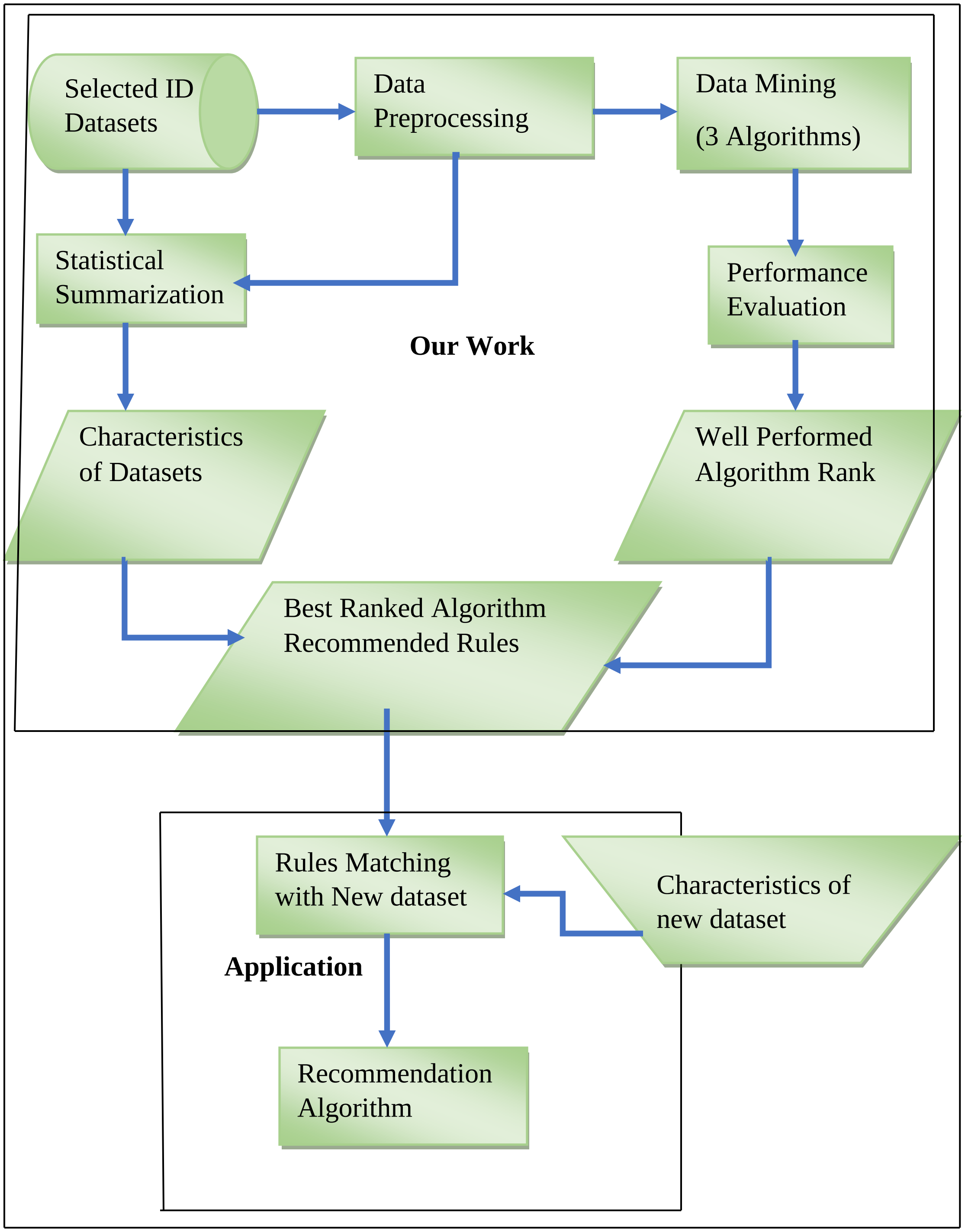

Selected ID Datasets
Data Preprocessing
Data Mining
(3 Algorithms)
Statistical Summarization
Performance Evaluation
Our Work
Characteristics of Datasets
Well Performed Algorithm Rank
Best Ranked Algorithm Recommended Rules
Rules Matching with New dataset
Characteristics of new dataset
Application
Recommendation Algorithm

Figure3. 2: General flow of the research method

3.5. Tools

The researcher is used a number of tools to complete the conducted research work. The following tools are used in the study:

✓ MS-Excel 2016: For encoding the paper data to make the data tabular for mining.

- ✓ Rapid Miner 7.4: For data preprocessing, Visualization, feature selection, Model building and evaluation of models.
- ✓ MS-Visual studio 2013: For developing graphical user interface.
- ✓ MS-Word 2016: For organizing the documentation
- ✓ Classification Algorithms such as Decision Tree (ID3 and J48), SVM (SMO), and ANN (Multilayer Perceptron): For developing the predictive model.

3.6. Algorithms Used in the Conducted Research

This research is conducted by adopting three data mining classification algorithms, which includes Decision Tree (DT) with Iterative Dichotomize3 (ID3) and C4.5 an extension of ID3 which is equivalent to J48 classifier, Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) for SVM classifier and Artificial Neural Network with Back Propagation Multilayer Perceptron classifier to develop the prediction model for the three diseases datasets. Classification is one of the major data mining tasks. So, this task is accomplished by generating a predictive model of data and interpreting the model regularly to provide information for selective labeled classes in data. Each algorithm used in this study is described in the following subsections.

3.6.1. Decision Tree Algorithm

Decision tree classification algorithm is one of the most widely used and practical methods for inductive inference. Decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost nodes in a tree are the root node. In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attribute values. The learned trees may also be represented with a set of IF-THEN rules to improve human readability (Jiawei al et, 2012).

Each leaf node is assigned to one class representing the most appropriate target value or the leaf node may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Internal nodes are represented with oval circles, whereas leaves are denoted by rectangles. It is used for classification of categorical variables and regression for continuous variables. This algorithm is highly used in medical domain for disease diagnosis, prediction, classification and prognosis (Jiawei al et, 2012). The following are the two decision tree algorithms, which is applied on the conducted research for the development of predictive models on the sample datasets.

3.6.1.1. Iterative Dichotomize3 (ID3) Classifier

ID3 is a predictive decision tree classification algorithm, which uses an information theoretic approach. The procedure is that at any point one examines the feature that provides the greatest gain information or equivalently and greatest decrease in entropy. For the general case in N labeled patterns partitioned into sets belonging to classes Ci, i = 1, 2, 3.... The population in class d is ni. Each pattern has n input features and each feature can take two or more values. The initial entropy values are calculated by using the following mathematical formula, where N is the total number of labeled patterns (Sushmita & Tinku, 2003).

Entropy=$\sum_{\mathrm{i}=1}^{\mathrm{l}} -\left(\frac{\mathrm{ni}}{\mathrm{N}}\right) \log_2 \left(\frac{\mathrm{ni}}{\mathrm{N}}\right) = \sum_{\mathrm{i}=1}^{\mathrm{l}} -\mathrm{p} \log_2$ pi---------------------Eq (4.1)

3.6.1.2. J48 Decision Tree Classifier

J48 classifier is a potential predictive classification algorithm, which uses gain ratio as splitting criteria. The splitting stops when the number of instances to be split is below a certain threshold. Error–based pruning is performed after the growing phase. It handles numeric attributes and encourages a training set that incorporates missing values by using corrected gain ratio criteria. The basic algorithm for J48 decision tree induction is a greedy algorithm which constructs a tree in a top down approach (dividing each node recursively until a leaf node is met). The following algorithm shows how J48 decision tree algorithm generates a tree from the given training data (Jiawei al et, 2012).

Input: The training samples, represented by discrete-value attribute; the set of candidate attributes, attribute-list.

Output: A decision tree.

Method

```
create a node N;
if tuples in D, are all of the same class, C then
return N as a leaf node labeled with the class C;
if attribute list is empty then
return N as a leaf node labeled with the majority class in D;
apply Attribute selection method (D, attribute list) to find the best splitting criterion;
label node N with splitting criterion;
if splitting attribute is discrete-valued and multiway splits allowed then
attribute list ← attribute list − splitting attribute; // remove splitting attribute
for each outcome j of splitting criterion // partition the tuples and grow subtrees for each partition
let Dj, be the set of data tuples in D satisfying outcome j; // a partition
if Dj, is empty then
attach a leaf labeled with the majority class in D to node N;
else attach the node returned by Generate decision tree (Dj, attribute list) to node N; end for
return N;
```

In this study J48 classifier was better in the classification of malaria and TB datasets than ANN and SVM. J48 classifier is, because the result is easy for interpretation and it gives the experiment result in the form of tree like structures. When take ANN or SVM the method of displaying results are highly dependent with different kinds of mathematical functions and needs additional interpretation experience to change the result in to analysis reports. Figure 4.6 and figure 4.7 shows the experiment results of J48 in cholera and malaria datasets in the tree like structure.

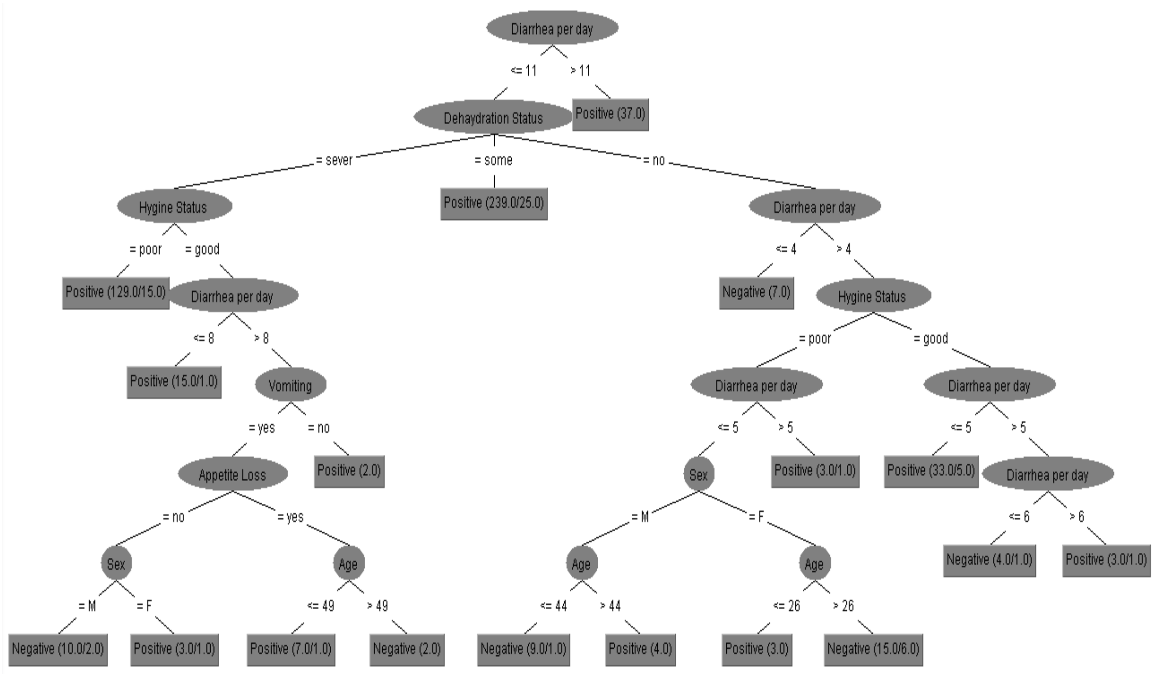


Figure3. 3:J48 classification results of cholera dataset using tree structure

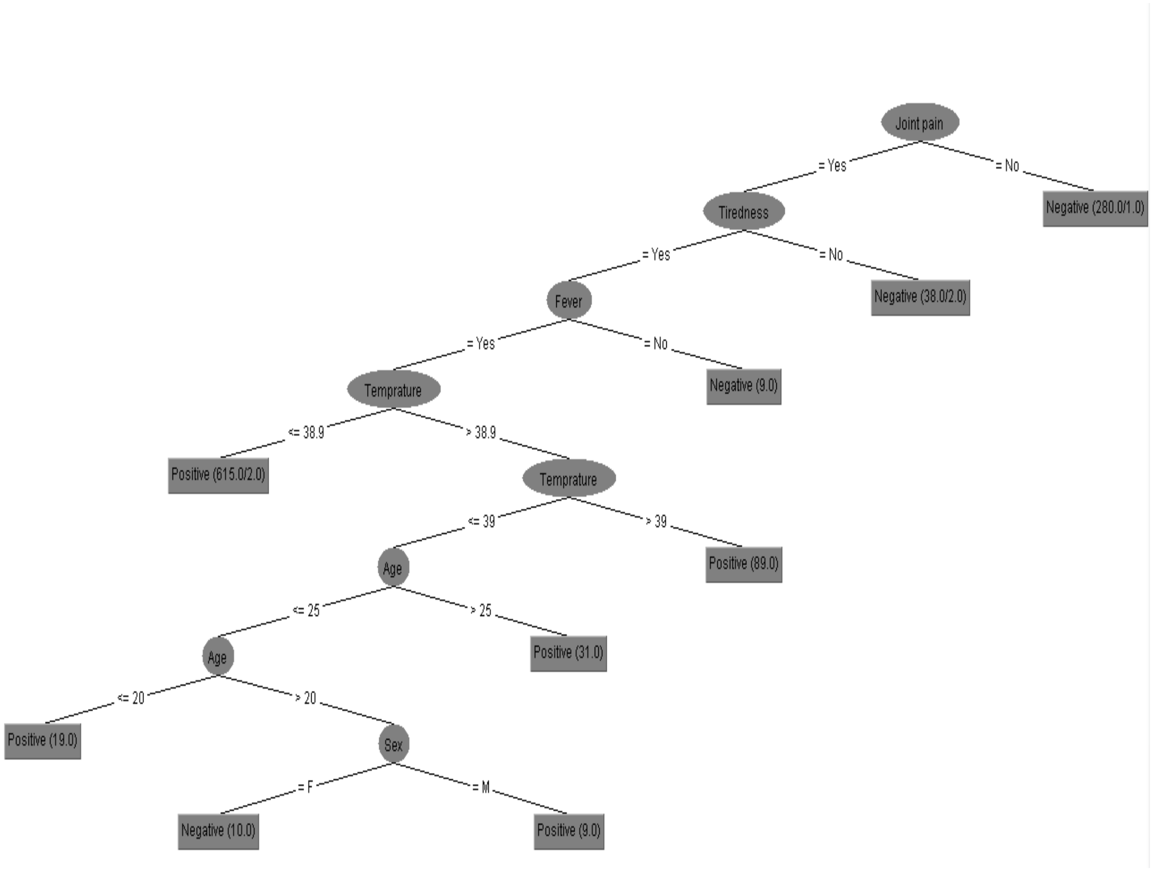


Figure3. 4:J48 classification result for malaria dataset using tree

3.6.2. Support Vector Machine (SVM) Classifier

Another classification algorithm used for developing the classification models is support vector machine (SVM). SVMs are supervised learning methods that generate input-output mapping functions from a set of labeled training data. The mapping function can be either a classification function used to categorize the input data or a regression function used for estimation of the desired output. For classification, nonlinear kernel functions are often used to transform the input data to a high dimensional feature space in which the input data becomes more separable compared to the original input space. Then, maximum-margin hyperplanes are constructed to optimally separate the classes in the training data. Two parallel hyperplanes are constructed on each side of the hyperplane, which separates the data by maximizing the distance between the two parallel hyperplanes. So larger the margin or distance between these parallel hyperplanes the better the generalization error rate of the classifier (David & Dursun , 2008).

SVMs are grouped with family of linear models, which works on classification or regression decisions based on the value of the linear combination of features. It shows highly competitive performance in many real-world applications, such as medical diagnosis, bioinformatics, face recognition, image processing and text mining. It is one of the most popular and state-of-the-art tools for knowledge discovery and data mining. Due to, its popularity in medical diagnosis and prognosis the researcher is used as one of the classification algorithm for developing the desired predictive models of the dataset.
In SVM the error rate of data mining is the sum of the training error rate and Vapnik Chervonenkis (VC) [1] dimension. In the given set of N training samples (Xi, yi), where Xi $\in R^n$ and yi ∈ {—1, 1}, then discriminate hyperplane is defined as

$$f(Xq) = \sum_{n=1}^{N} yi\alpha i \, K(Xq, Xi) + b$$------------------------------------Eq (4.2)

Here K (.) is a kernel function and the sign of f(Xq)determines the membership of query sample Xq. Constructing an optimal hyperplane is equivalent to determining all nonzero αi, which correspond to the support vectors, and the bias b(Sushmita & Tinku, 2003).Figure3.2 showed the process of developing a model using SVM classification algorithm adopted from Sushmita& Tinku.

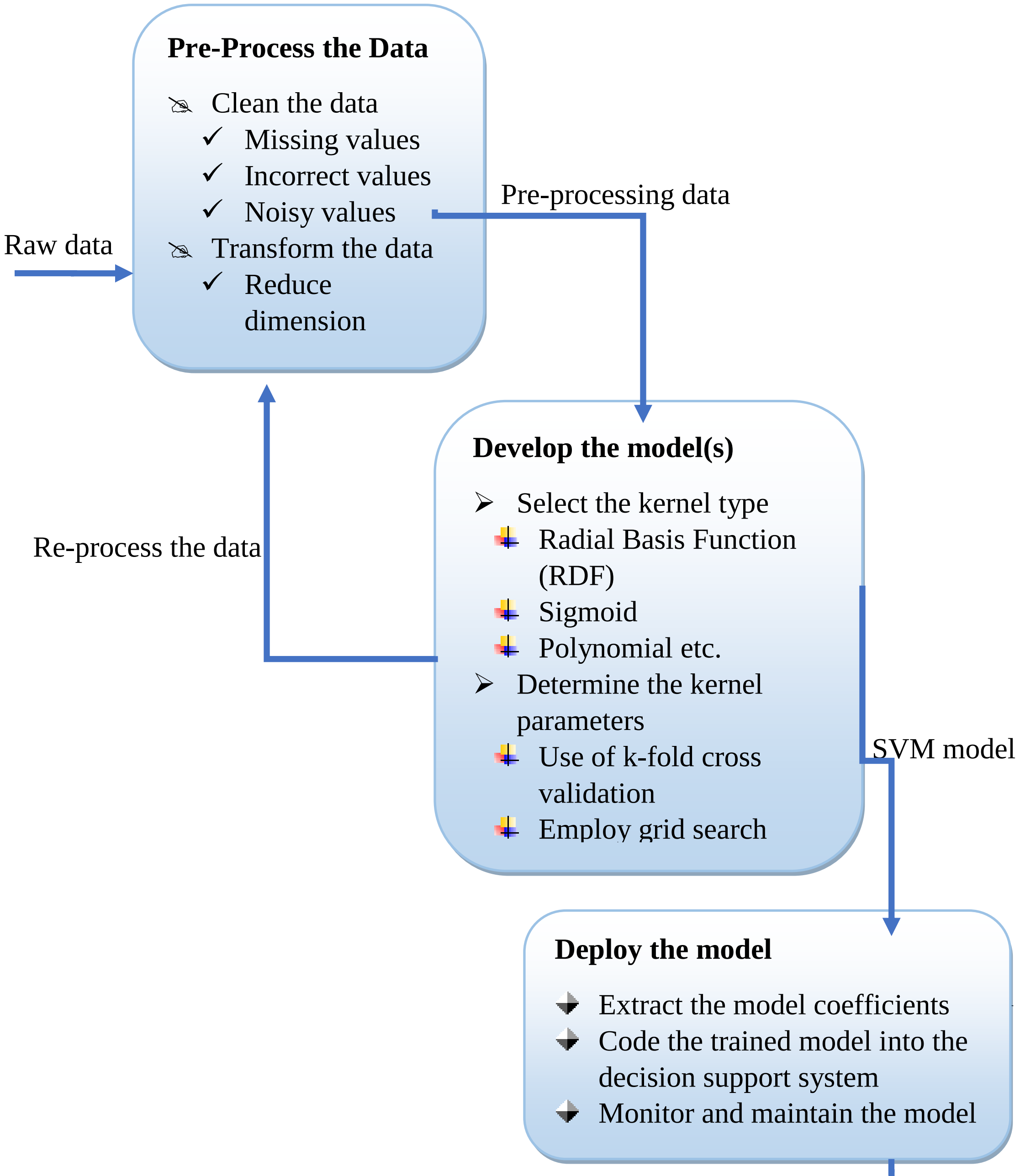


Figure3. 5: Process description for SVM model development

3.6.3. Artificial Neural Network

ANN is a mathematical representation of human neural architecture; it reflects learning and simplification abilities. One of the unique aims of artificial neural networks (ANN) is understand and shape the functional characteristics and computational properties of the human brain when it performs cognitive processes such as sensor perception, concept categorization, concept association and learning. ANN has at least two physical components such as the processing elements and the connections between them. The processing elements are neurons and the connections are links. Every link has a weight parameter associated with it and each neuron receives stimulus from the neighboring neurons connected to it, processes the information and produces an output. Neurons that receive stimuli from outside the network are input neurons, neurons whose outputs are used extremely are output neurons and Neurons that receive stimuli from other neurons and whose output is a stimuli for other neurons are hidden neurons. A number of ANN algorithms are widely applied in the fields of medical science around the world. The back propagation multilayer Perceptron algorithm is the algorithm, which is considered in this study since it is the most widely used algorithm in the study of neural networks. Often, back propagation algorithm comprises the following three steps (Amato et al., 2013).

- The input pattern is presented to the network where by the input pattern is propagated through the network until they reach the output units. This forward pass would produce the actual or predicted output.
- Then the desired output would be given as part of the training set, so that the actual output can be subtracted from the desired output in order to give the error signal.
- In the last step the errors are passed back through the neural network by computing the contribution of each hidden processing unit and deriving the corresponding adjustment needed to produce the correct output. The connection weights are then adjusted and the neural network is said to have learned from an experience. The above steps are repeatedly carried out for all samples in the data until the termination condition of the iteration is satisfied.

A Multilayer Perceptron (MLP) is the most known and most frequently used type of neural network. It is a special type of feed-forward network employing three or more layers, with nonlinear transfer functions in the hidden layer neurons. The signals are transmitted within the network in one direction: from input to output. That is, there is no loop; the output of each neuron does not affect the neuron itself. This architecture is called feed forward neural network. In the MLP structure, the neurons are grouped into layers. The first and last layers are called input and output layers respectively because they represent inputs and outputs of the overall network. The remaining layers are known as hidden layers. It is particularly suitable for medical diagnosis applications where the inputs and outputs are numerical and pairs of input/output vectors are providing a clear basis for training in a supervised manner (Jiawei al et, 2012).

Typically, a multilayer Perceptron consists of a set of sensory units or source nodes that constitute the input layer, one or more hidden layers of computational nodes and an output layer of computational nodes. The input signal propagates through the network in a forward direction on a layer-by-layer basis. Multilayer Perceptron have been applied successfully to solve difficult and diverse problems by training them in a supervised manner. It iteratively learns a set of weights for prediction of the class label of tuples (Jiawei al et, 2012).

The inputs to the network correspond to the attributes measured for each training tuples. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The numbers of hidden layers are arbitrary. The weighted outputs of the last hidden layer are input to units making up the output layer. The units in the input layer are called input units. The units in the hidden layers and output layer are sometimes referred to as neurodes, due to their symbolic biological basis, or output units.

Three layers were used in the study to develop the network architecture. The first layer has an input layer, which contains attributes of each sample dataset, and the outputs of different input layer attributes are become an input for the next layer called the first hidden layer and the hidden layers are gives two different outputs called positive and negative values. As because feed forward multilayer Perceptron network with one or two hidden layers are able to learn everything and this deep learning facilitates a more complex representation of patterns in the data only single hidden layer was used to develop the classification models of the data in the study. A feed forward network with one or two hidden layers can function virtually as a universal approximation for any mathematical function and also the network with one hidden layer has two processing layers such as the hidden layer and output layer. Another reason why a single hidden layer is used in the study is making the number of hidden layers are two or more than two, it require a very extensive training set to be able to compute weights for the network. Deep learning in feed forward network means networks of several hidden layers. Figure 3.4 below shows the network architectures of the malaria dataset in the study, which is used the development of the classification model. The structure of the network is also the same for other sample datasets the only difference is the types of input variables, hidden layer, weighted values for each neuron, and types of hidden layers produced the final output layers they used. There are five hidden neurons in the hidden layers of the network and each has different weighed values.

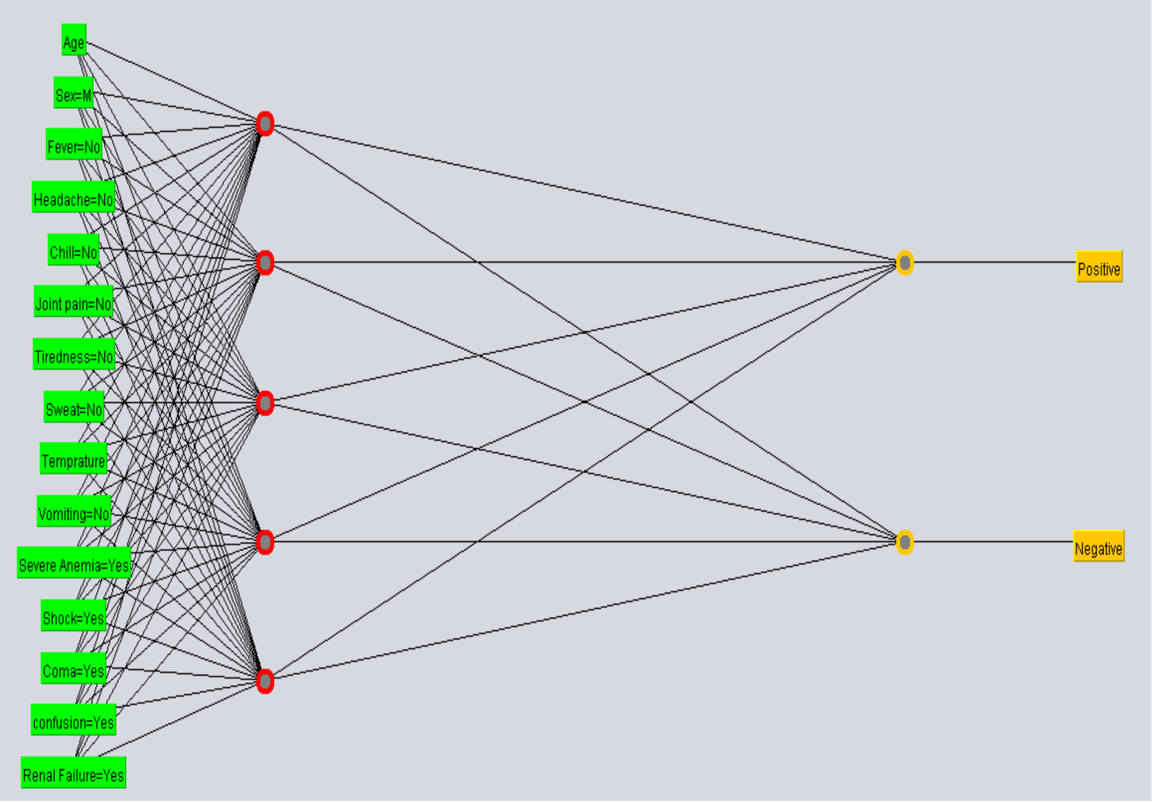


Figure3. 6: MLP Network Structure for Malaria Dataset

CHAPTER FOUR

4. Data Processing, Model Building and Performance Evaluation

4.1. Data Format

Once the researcher is decided to use Weka 3.8 and Rapid miner 7.4 data mining software tools for the development of the required predictive model, the next step is making the available data format suitable for Weka and Rapid miner.

The first thing we do is encoding the patients' medical history from paper format into an excel file format (.xls) by identifying all variables important for each dataset document. Then the sample data sets are saved separately in CSV (Comma Separated Values) sometimes called Comma Delimited File format. The CSV file format is a specially formatted plain text file often used as a simple way to transfer a large volume of spreadsheet or database information between programs, without worrying about special file types and both Weka and Rapid miner are read CSV file formats. But Weka further reads ARFF (Attribute-Relation File Format) file formats. So, CSV files are converted into ARFF format, which is the default file format for Weka but is not suitable for rapid miner. An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes.

Each record in the sample dataset corresponds to a single patient's medical history, which contains demographic information, clinical symptoms, laboratory examinations and results, medical treatments and prescriptions ordered by physicians who are collected during inpatient or outpatient medical services provided by hospitals. From this recorded information, the researcher has selected attributes, which helps to develop the predictive model for the diagnosis and treatment of infectious disease patients.

Initially the study contains demographic attributes such as Card No, Unit No, First name, Middle name, Last name, Address, Age and Sex for the three dataset samples, and Medical information's like Lab prescriptions, X-ray prescription, Laboratory results, X-ray results and clinical symptoms such as Watery diarrhea, Diarrhea loss Per day, Vomiting, Hygiene status, Weight loss, Appetite lose, Dehydration status and Cholera

Status (class variable) for cholera (Acute Watery Diarrhea) disease, Fever, Headache, Chill, Joint pain, Tiredness, Temperature, Vomiting, Sweating, Sever anemia, Shock, Coma, Confusion, Renal failure and Malaria Status (Class variable) for Malaria and Weight, HIV test performed, HIV test result, Headache, Cough(>2wks), Chest pain, Fever, Weight loss, Loss of appetite, Shortness of breath, Productive sputum, Night sweating and TB Status(Class variable) for TB. Demographic variables like name, address, marital status, and family history, and clinical information's like type of examination, types of medicine prescribe by physicians and date of service were not important to develop the model, so the researcher was not included the above variables in the study dataset. As a result the study has initially contains a total of 18 attributes for cholera, 24 attributes for malaria and 25 attributes for TB disease. However from this number 10, 16 and 15 attributes for cholera, malaria and TB respectively are selected for developing the intended model. From this best selected attributes 7, 10 and 10 attributes for cholera, malaria and TB respectively are also selected using attribute subset selection technique to check the performances of the classifiers with limited and large number of attributes. The next three tables are shows the number of selected attributes from the total number of attributes included in the dataset and their detail descriptions.

Table4. 1: Cholera (AWD) Disease Attributes and descriptions

No	Attribute Name	Attribute Description	Type
1	Age	Age of the patient in years	Numeric
2	Sex	Sex of a patient with value (Male, Female)	Nominal
3	Watery diarrhea	Whether the patient has diarrhea (Yes, No)	Nominal
4	Diarrhea per day	How much time a patient loses diarrhea per day	Numeric
5	Vomiting	Have a patient vomits (Yes, No)	Nominal
6	Weight loss	Does a patient lose physical weights (Yes, No)	Nominal
7	Appetite loses	Whether a patient loses interest to eat (Yes, No)	Nominal
8	Dehydration status	At what level, a patient has losing dehydration (some, sever, never)	Nominal
9	Hygiene status	A patient living condition (Good, Poor)	Nominal
10	Cholera Status	Patients cholera test result (Positive, Negative)	Nominal

Table4. 2: Malaria Disease attributes and descriptions

No	Attribute name	Attribute Description	Type
1	Age	Age of a patient in year	Numeric
2	Sex	Sex of a patient with value (Male, Female)	Nominal
3	Fever	Does a patient have fever or body heat (Yes, No)	Nominal
4	Headache	Does a patient have a headache (Yes, No)	Nominal
5	Chill	Does a patient have body freezing (Yes, No)	Nominal
6	Joint pain	Does a patient have pain on muscles (Yes, No)	Nominal
7	Tiredness	Does a patient have weakness or fatigue (Yes, No)	Nominal
8	Temperature	A patient's body temperature on degree ℃	Numeric
9	Vomiting	Does a patient have sickness or vomiting (Yes, No)	Nominal
10	Sweating	Does a patient have sweat from the body (Yes, No)	Nominal
11	Severe anemia	Whether a patient has anemia (Yes, No)	Nominal
12	Shock	Does a patient has shocked or tremor (Yes, No)	Nominal
13	Coma	Does a patient has in high risk or coma (Yes, No)	Nominal
14	Confusion	Does a patient has confused in things (Yes, No)	Nominal
15	Renal failure	Does a patient have failure of them self (Yes, No)	Nominal
16	Malaria Status	Patients malaria test result (Positive, Negative)	Nominal

Table4. 3: Tuberculosis Disease attributes with its descriptions

No	Attribute Name	Attribute Description	Type
1	Age	Age of patient in numeric year	Numeric
2	Sex	Sex of patient with value (Male, Female)	Nominal
3	Weight	Physical weight of patient in number	Numeric
4	HIV Performed	Does a patient has tested for HIV (Yes, No)	Nominal
5	HIV Test Result	Patient test result for HIV (R, NR, Not)	Nominal
6	Headache	Does a patient has a headache or pain (Yes, No)	Nominal
7	Cough (>2wks)	Does a patient coughed above 2wks. (Yes, No)	Nominal
8	Chest pain	Does a patient has pain on the chest (Yes, No)	Nominal
9	Fever	Increase in temperature from normal state (Yes,	Nominal

		No)	
10	Weight loss	Does a patient reduced weights (Yes, No)	Nominal
11	Appetite Loss	Does a patient loses interest to eat (Yes, No)	Nominal
12	Shortness of Breath	Does a patient has a breathing problem (Yes, No)	Nominal
13	Productive Sputum	Does the patient have sputum (Yes, No)	Nominal
14	Night Sweating	Whether a patient sweat at night (Yes, No)	Nominal
15	TB Status	Patients test result (Positive, Negative)	Nominal

4.2. Data Preprocessing

As noted by (Kariuki et al , 2016), today's real-world databases are highly vulnerable to noisy, missing, and inconsistent data, which is typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results. Incomplete, noisy, and inconsistent data are commonplace properties of large real-world databases and data warehouses (Padma , 2004). Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, other data may not be included simply because it was not considered important at the time of entry, relevant data may not be recorded due to misunderstanding, or equipment malfunctions, data that were inconsistent with other recorded data may have deleted and missing data, particularly for tables with missing values for some attributes, may need to be inferred (Han and Kamber, 2006). A number of data preprocessing techniques were applied on the data, such as Data cleaning to handle missing values, remove noise, to handle outliers and correct inconsistencies in the data, data transformation, dimensionality reduction and feature selection have been done. Figure4.1 showed the data preprocessing tasks for quality data mining (Padma, 2004).

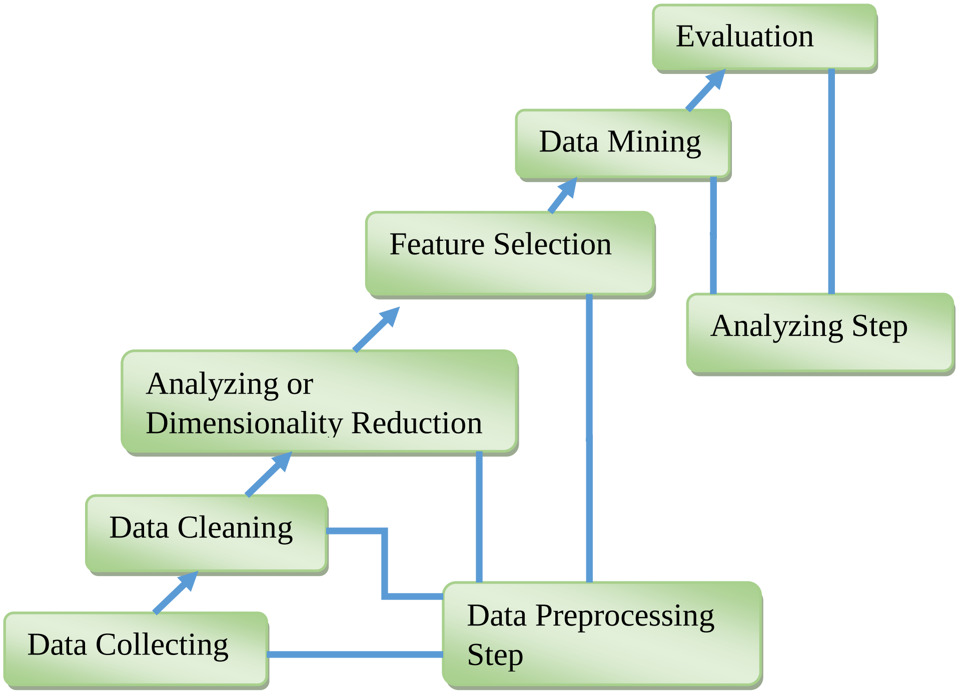


Figure4. 1: Data Preprocessing Tasks for Quality Data Mining

4.2.1. Data Cleaning

Data cleaning can be applied to remove noisy and correct inconsistencies in the data. Quality of data plays important roles in information oriented organizations like healthcare, where the knowledge is extracted from data. Consistency, completeness, accuracy, validity and timeliness are important characteristics of quality data (Divya & Sonali , 2014). Data cleaning is a process, which involves filling in missing values, smooth noisy, identify and remove outliers, resolving inconsistencies and improve the quality of data to be mined. In this study filling missing values and removing noise data is done in the data cleaning stage of the study.

i. Handling Missing Values

Missing values or incomplete data are becoming serious problems in many fields of research**.** Missing data may be recorded because, the data may not be entered due to misunderstanding, the data which is inconsistent with other recorded data, the deleted data may not be considered as important at the time of entry and the data changes have not been recorded. It is handled by various ways like ignoring the records when an entered tuples are empty, filling the missing values manually, using a global constant, which fills the same value for all missing attributes, attribute mean value, media or most probable value methods (Padma, 2004). Identifying and solving those missed values is needed to prepare the dataset for experimentation. The researcher is used rapid miner 7.4 mining tool to identify and select an attribute which contains an incomplete value and the number of missing values for each attribute with exploratory analysis and took necessary measure to handle missing values.

In the total sample datasets of the study, attributes like Age, Sex, Headache and chest pain was consisting 1, 2, 3, and 3 missing values, respectively, for TB case, Age, Sweat and vomiting have consisting 2, 4, and 1 missing values respectively, for malaria case and weight loss and appetite lose have consisting 1 missing values each for cholera case due to missing of values by recorders at the time of encoding the data from paper to excel format. To handle the missing records, the researcher uses the attribute mean value method for numeric variable Age in all cases. However, for all other attributes, which contains missing values the researcher is used most probable value, replacing technique to fill the missing value because using other techniques for nominal attributes are not convenient.

ii. Remove Noisy

Noisy data are due to random errors or changes in a measured variable. There may be incorrect attribute values in the data due to, faulty data collection instruments, data entry problems, data transmission problems, duplicate records, incomplete and inconsistent data. In the study sample datasets, an attribute weight in TB case consisting of 5 negative values -48kg, -65kg, -16kg, -40kg; -32kg and 2 rational values 0.56kg, 0.52kg due to encoder error. Thus, the researcher has discussed with domain experts and data encoder experts. Finally, the researcher removes the inconsistent and noise attribute values by using the correct numeric attribute value 48kg, 65kg, 16kg, 40kg, 32kg and 56kg, 52kg respectively to make the data consistent for mining.

4.2.2. Data Transformation

Data transformation is about transforming or consolidating the data to make it appropriate for mining. This involves adding derived fields to bring information to the surface. It may include smoothing, aggregation, generalization, normalization, Discretization, and attribute construction. To make the dataset suitable for this study data Discretization technique is applied on numeric attributes to minimize distinct values of attributes, dimensionality reduction is also used to reduce the size of the dataset and attribute selection method is the last method applied to remove weakly relevant attributes.

4.2.2.1. Data Discretization

Binning method is used to reduce the size of data by dividing the range of numeric value attributes into interval continuous value attributes. Divide the range of numeric values into N intervals of equal size and give label for each interval. Attributes such as age and physical weight from TB, age and body Temperature from malaria, and age and diarrhea per day from cholera cases are discretized into six equal intervals binning to make the continuous value attributes are valuable for mining purpose. The following three tables are shown the Discretized value attributes in this study. All the continuous variables in each case are discretized separately.

Table4. 4: Discretized attributes of Age and Physical Weight for TB patients

Age (Binned) Label (year)	Frequency	Percentage %	Weight (Binned) Label (kg)	Frequency	Percentage %
(1-14]	161	**12**	(1-16]	67	**5**
(14-27]	516	**37**	(16-28]	78	**5**
(27-41]	405	**29**	(28-39]	92	**7**
(41-54]	144	**10**	(39-51]	444	**32**
(54-67]	89	**6**	(51-63]	534	**38**
(>67)	77	**6**	(>63)	177	**13**

Table4. 5: Discretized attributes of Age and Body Temperature for Malaria patients

Age Binned Label(year)	Frequency	Percentage %	Body Temperature Binned Label (℃)	Frequency	Percentage %
(1-15]	164	**15**	(0-37]	182	**17**
(15-25]	518	**47**	(37-38]	542	**49**
(25-36]	261	**24**	(38-39]	295	**27**
(36-46]	108	**10**	(39-40]	46	**4**
(46-57]	36	**3**	(40-42]	32	**3**
(>57)	13	**1**	(>42)	3	**0**

Table4. 6: Discretized attributes of Age and Diarrhea Loss per Day for Cholera patients

Age Binned Label(year)	Frequency	Percentage %	Diarrhea Loss per Day Binned Label (Number)	Frequency	Percentage %
(1-20]	83	**16**	(1-5]	144	**27**
(20-31]	100	**19**	(5-7]	150	**29**
(31-43]	127	**24**	(7-9]	129	**25**
(43-55]	105	**20**	(9-11]	65	**12**
(55-66]	72	**14**	(11-13]	30	**6**
(>66)	38	**7**	(>13)	7	**1**

4.2.2.2. Attribute Selection

Attribute or feature selection is a part of data transformation with dimension reduction. Its objective is to identify attributes in the dataset which is highly important for developing the predictive model and discard any other features that is irrelevant for the model rather only provides redundant information and maximize the size and dimensionality of the dataset.

There are so many potential benefits for variable and feature selection in data mining. This includes facilitating data visualization and data understanding, reduce data dimensions and storage requirements, reduce training and operation times, enables learning algorithms to operate faster and more effectively and finally challenging the curse of dimensionality to improve prediction performance (Guyon & Andre , 2003).

Feature selection algorithms have two principal components. These are a selection algorithm that generates proposed subsets of features and attempts to find an optimal subset and an evaluation algorithm that determines the goodness of a proposed feature subset in respect to the selection algorithm. Feature selection methods to search through the subsets of features and try to find the best one among the competing 2^N candidate subsets according to some evaluation function (Kotsiantis et al, 2006).

Feature Selection methods are grouped into three broad categories, i.e. filter, wrapper and embedded approaches based on their requirement on the inductive algorithm that will finally use the selected subset. Filter methods operate independently of the learning algorithms while wrapper methods take into account the learning algorithms to be used.

To select the top ranked attributes from the total list of attributes in each sample dataset, the researcher is used both filter and wrapper feature selection approaches by applying Gain ratio attribute evaluator to the ranker search method. Gain ratio attribute evaluator evaluates the worth of an attribute by measuring the gain ratio with respect to the class and rank all attributes top to bottom based on gain value. Based on this the researcher has selected 7 top ranked attributes from a total of 10 attributes for cholera, 10 top ranked attributes from a total of 16 attributes for malaria and 10 best ranked attributes for TB from a total of 15 attributes including class attributes for all cases. Table4.7. shows all the selected attributes of each case using gain ratio attribute evaluator with filter and wrapper features selection approaches.

Table4. 7: Top ranked attributes selected for model development

Cholera Case		**Malaria Case**		**TB Case**	
No	**Attribute Name**	**No**	**Attribute Name**	**No**	**Attribute Name**
1	Watery Diarrhea	1	Joint Pain	1	Productive Sputum
2	Dehydration status	2	Chill	2	Weight loss
3	Age	3	Tiredness	3	Chest Pain
4	Diarrhea per day	4	Fever	4	Night Sweating
5	Vomiting	5	Headache	5	HIV Test Result
6	Weight Loss	6	Severe Anemia	6	Cough(>2wks)
7	Cholera Status	7	Renal Failure	7	Fever
		8	Age	8	Age
		9	Shock	9	Shortness of Breath
		10	Malaria Status	10	TB Status

After selecting attributes from each case, the investigator, compares the prediction ability of selected attributes with respect to all attributes in different classifiers used for prediction in the conducted research work.

4.3. Data mining and Model Development

In the previous subsections tasks appropriate for developing the required classification models are completed. These include collecting sample datasets from the problem area, which is used for training and model development, identifying the mining process model, which is important for the selected problem, cleaning and organizing the data such as performing data treatments like replacing missing values, removing outliers, smoothing inconsistencies and preparing a complete dataset, making a descriptive analysis of the data with statistical distributions of attributes in the datasets and making the data useful for model-building such as transforming with Discretization and selecting attributes highly important for the modeling activity. The next main activity for the study is developing the classification model using classification algorithms, Decision Tree (ID3 and J48), Support Vector Machine (SMO) and Artificial Neural Network (MLP) classifiers and evaluates the performance of each model in the three different disease datasets such as cholera (AWD), malaria and tuberculosis (TB).

4.4. Experimentation

This section presents the steps and procedures followed during the experimentations for model development to classify cholera, malaria and TB patients based on the historical data taken from hospital data stores. The main objective of this study is, discovering new hidden patterns for predicting occurrences of cholera, malaria or TB in the patient's body. Having this purpose in mind, the researcher has done model-building, which is carried out by using classification data mining approach. As the study has conducted with an experimental approach, a total of 24 different models are developed from four main experiments with each experiment has six separate classification models generated by using Decision tree with ID3 and J48 classifier, Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) for SVM classifier and Artificial Neural Network with Back Propagation Multilayer Perceptions (MLP) classification algorithms in the three different datasets.

After preprocessing is done on the data a total of 3,017 instances are ready for experimentation. The data have partitioned into three dataset classes based on the name of the disease in which samples are taken. Those are 525 instances are represented cholera, 1100 instances are represented malaria and 1392 instances are represented TB patients. The model development in this investigation is carried out in two phases. In the first phase of this experiment the researcher is planned to check the performance of the four selected algorithms in the three different datasets using all attributes of each dataset and in the second phase the researcher also checks the performance of the above algorithms by applying attribute subset selection which is filtered by Gain ratio attribute evaluator with ranker search method to identify effects of attribute selection on the performances of a particular mining algorithm.

The test option in this study is 10-fold cross validation, which partitions the total sample dataset into training and test set. At experimentation the sample data is divided into 10-

folds by the tool. From 10-folds, 9-folds are assigned for training and the remaining (1-fold) 10th-fold is intended for testing the model, then its error rate is calculated on the holdout set. Therefore, the learning process is performed in a total of 10 times on each experiment (Witten & Frank, 2005). The researcher decided to use 10- fold test option due to its ability to perform widespread tests on many datasets with different learning techniques and 10 is the right number of folds to get the best estimate of error and there is also some theoretical evidence that backs this up. The conducted experiments on each data mining algorithm with all attributes and filtered attributes in the three sample datasets are discussed next.

4.4.1. Model Development via ID3 Decision Tree Classifier

In this experiment, the performance of ID3 classifier is evaluated on three different datasets, which contains different instances by applying all attributes. Six predictive models are built with default parameters using all 10, 16, 15 and best selected 7, 10, and 10 attributes of cholera, malaria and TB datasets respectively. Table 4.8 shows the predictive ability of ID3 in the three datasets. Its predictive performance is measured by TP and FP rates, Precision, Recall, F-Measure, ROC Area and Accuracy.

Table4. 8:ID3 classification results using all and best selected attributes

Model	Number of instances	Time (Sec)	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy
ID3 with all (10) attributes applied to cholera dataset	525	0.03	0.824	0.603	0.799	0.824	0.808	0.672	81.14%
ID3 with all (16) attributes applied to malaria dataset	1100	0.03	0.984	0.029	0.984	0.984	0.984	0.988	98%
ID3 with all (15) attributes applied to TB dataset	1392	0	1.00	0.00	1.00	1.00	1.00	1.00	99.2%
ID3 with best 7 attributes applied to cholera dataset	525	0	0.805	0.746	0.752	0.805	0.771	0.709	78.48%
ID3 with best 10 attributes applied to malaria dataset	1100	0	0.985	0.028	0.986	0.985	0.985	0.979	98.45%
ID3 with best 10 attributes applied to TB dataset	1392	0	0.992	0.011	0.992	0.992	0.992	0.992	99.2%

From the experiment performance of ID3 classifier is evaluated in different dataset sizes using all and best selected attributes of each dataset and using 10-fold cross validation evaluator. For Example, in the Cholera dataset out of 525 instances only 426 (81.14%) instances are correctly classified the remaining 91 (17.33%) instances are incorrectly classified in positive or negative status and 8 (1.52%) instances are unclassified by ID3 when applying all attributes for model development. Whereas only best selected 7 attributes are applied to model development only 412 (78.48%) instances are correctly classified, 100 (19.05%) instances are incorrectly classified in different cholera status classes and 13 (2.48%) instances are also unclassified in any class by ID3 classifier. The overall accuracy of ID3 in cholera dataset is 81.14%, the model developed by applying all attributes in the algorithm. On the other hand, in the malaria dataset out of 1100 instances 1078(98%) instances are correctly classified by the classifier and only 18(1.64%) instances are incorrectly classified and 4(0.36%) instances are unclassified in any class by applying all attributes and out of 1100 malaria cases 1083(98.45%) instances are correctly classified, 16(1.45%) instances are incorrectly classified and only 1(0.09%) instance is unclassified by ID3 classifier by using best selected attributes for developing the model. The overall predictive accuracy of ID3 in malaria dataset is 98.45% by applying best selected attributes of the dataset. In the TB dataset ID3 is showed high accuracy than the other two datasets. Based on this out of 1392 TB cases 1381(99.2%) instances are correctly classified, 10(0.72) % instances are incorrectly classified) and only 1(0.07%) instance is unclassified by ID3. The overall accuracy of ID3 in TB dataset is 99.2% both in 15 and best selected 10 attributes.

Based on the above experiment attribute subset selection decreases accuracy, TP rate (sensitivity), Precision, Recall and F-measure, but increased FP rate (specificity) and ROC area in the cholera dataset and attribute subset selection increases accuracy, TP rate, Precision, Recall and F-measure and decreased FP rate and ROC areas in malaria dataset and have similar performances whether attribute subset selection is applied or not in the TB dataset.

4.4.2. Model Development via J48 Decision Tree Classifier

In this experiment, the predictive ability of J48 classifier is investigated using the three patient datasets that contain different number of instances. In this experiment six separate models are built with unpruned Tree by applying all 10, 16, and 15 attributes as well as best selected 7, 10 and 10 attributes of cholera, malaria and TB patient data sets respectively. Table 4.9 shows the experimental results of J48 classifier with all attributes.

Table4. 9:J48 classification results by applying all and best selected attributes

Model	Number of instances	Time (Sec)	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy
J48 with all (10) attributes applied to cholera dataset	525	0.03	0.827	0.624	0.798	0.827	0.807	0.663	82.66%
J48 with all (16) attributes applied to malaria dataset	1100	0	0.985	0.028	0.986	0.985	0.985	0.970	98.55%
J48 with all (15) attributes applied to TB dataset	1392	0	0.992	0.011	0.992	0.992	0.992	0.992	99.2%
J48 with 7 best attributes applied to cholera dataset	525	0	0.811	0.719	0.764	0.811	0.779	0.668	81.14%
J48 with 10 best attributes applied to malaria dataset	1100	0	0.985	0.028	0.986	0.985	0.985	0.970	98.55%
J48 with 10 best attributes applied to TB dataset	1392	0	0.992	0.011	0.992	0.992	0.992	0.992	99.2%

From table 4.9, we have seen accuracy, Precision, Recall, TP rate (Sensitivity) and F-measures of J48 classifier have maximum in TB patient dataset and minimum in the cholera dataset. This refers the prediction of J48 classifier highly depends on the features and sizes of a particular dataset. For example, from 525 instances of cholera patients only 434 (82.67%) instances are correctly predicted by J48, but the remaining 91 (17.33%) instances are incorrectly or falsely predicted by J48 classifier by applying all attributes in the classifier and out of 525 instances of cholera only 426 (81.14%) are correctly predicted but the remaining 99 (18.86%) instances are incorrectly predicted. This shows additional 8 instances are incorrectly predicted by J48 classifier when applying best selected attributes of the dataset in the classifier. On the other hand, from 1100 instances of malaria patients 1084 (98.55%) instances have correctly predicted as a malaria case and the remaining 16 (1.45%) instances are incorrectly predicted by the classifier when applying either all or best selected attributes. Whereas the classification of J48 is 99.2% of TB cases and only 0.8% of instances are incorrectly predicted by the classifier by applying either all or best selected attributes.

From the above experiment, we have shown that applying best selected attributes in the classifier does not improve the performance of J48 using unpruned prediction technique in the three datasets. Rather, its accuracy is reduced from 82.67% to 81.14% in cholera dataset and has the same predictive accuracy for other datasets. This refers attribute subset selection with 10-fold cross validation evaluation technique may not improve the performance of J48 classifier.

4.4.3. Model Development via Support Vector Machine Classifier

The main purpose of this experiment is evaluating the performances of the SVM classifier by applying Sequential Minimal Optimization (SMO) algorithm by using Radial Basis Function Network (RBF) classifier for calibrator and RBF Kernel as a kernel property with all and best selected attributes that have selected using gain ratio attribute subset selection technique in the three patient datasets. In this experiment six separate models are built using the whole features of the given datasets and the changed properties of SMO. Table 4.10 shows the prediction power of the SVM classifier within the three patient datasets using all 10, 16 and 15 attributes or best selected 7, 10 and 10 attributes of cholera, malaria and tuberculosis respectively.

Table4. 10: SVM classification results by applying all and best selected attributes

Model	Number of instances	Time (Sec)	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy
SVM with all (10) attributes applied to cholera dataset	525	0.06	0.834	0.834	0.696	0.834	0.759	0.500	83.43%
SVM with all (16) attributes applied to malaria dataset	1100	0.22	0.973	0.060	0.974	0.973	0.972	0.956	97.27%
SVM with all (15) attributes applied to TB dataset	1392	0.17	0.992	0.011	0.992	0.992	0.992	0.992	99.2%
SVM with all (7) attributes applied to cholera dataset	525	0.05	0.834	0.834	0.696	0.834	0.759	0.500	83.43%
SVM with all (10) attributes applied to malaria dataset	1100	0.14	0.969	0.069	0.970	0.969	0.969	0.950	96.91%
SVM with all (10) attributes applied to TB dataset	1392	0.13	0.992	0.011	0.992	0.992	0.992	0.992	99.2%

The experiment results tell us, the prediction accuracy of SVM is higher in TB dataset, but minimum prediction in cholera dataset. For example, from 525 instances of cholera cases only 438 (83.43%) are correctly predicted by SVM but the remaining 87 (16.57%) instances are incorrectly predicted by SVM. When we see the TB dataset, 1382 (99.2%) instances are correctly predicted by SVM classifier. Accuracy of SVM is also higher in malaria dataset, out of 1100 sample instances 1070 (97.27%) are correctly predicted but have a less prediction accuracy compared to J48 and ID3 classifiers.

Even though, attribute subset selection speeds up the learning time to develop the prediction model. Its prediction accuracy is decreased or remained the same in both methods. For example, applying all attributes for model development takes 0.06, 0.22 and 0.17 seconds for learning to build model for cholera, malaria and TB cases respectively. But this is reduced to 0.05, 0.14 and 0.13 seconds for developing the model in cholera, malaria and TB datasets respectively after applying best filtered attributes on the SVM classifier. These shows applying best filtered attributes for model development on SVM classifier does not improve its classification accuracy rather it minimize time for developing models.

4.4.4. Model Development via Artificial Neural Network classifier

This experiment aims to evaluate the performance of ANN using the multilayer Perceptron (MLP) with back propagation classifier in the three dataset samples. In this investigation, multilayer Perceptron classifier with a back propagation method using default parameters is applied on the datasets. In the experiment six separate models are built by applying all and best filtered attributes of each dataset and show its prediction in different attribute variables. Table 4.11 shows the performance of multilayer Perceptron in the three datasets with or without attribute subset selection is done on the classifier.

Table4. 11: MLP classification results by applying all and best filtered attributes

Model	Number of instances	Time (Sec)	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy
MLP with all (10) attributes applied to cholera dataset	525	1.19	0.802	0.574	0.791	0.802	0.796	0.752	80.19%
MLP with all (16) attributes applied to malaria dataset	1100	3.16	0.985	0.028	0.985	0.985	0.984	0.991	98.45%
MLP with all (15) attributes applied to TB dataset	1392	4.36	0.992	0.011	0.992	0.992	0.992	0.992	99.2%
MLP with all (7) attributes applied to cholera dataset	525	0.9	0.796	0.731	0.749	0.796	0.768	0.748	79.62%
MLP with all (10) attributes applied to malaria dataset	1100	1.35	0.985	0.028	0.986	0.985	0.985	0.983	98.55%
MLP with all (10) attributes applied to TB dataset	1392	2.08	0.992	0.011	0.992	0.992	0.992	0.992	99.2%

From the result in the above experiment prediction accuracy of MLP classifier is higher in TB cases, which is 99.2% and has lower classification on cholera cases, which is 80.19%. For example, out of 525 samples of cholera patients only 421 (80.19%) instances are correctly classified by MLP classifier, but the other 104 (19.81%) instances are incorrectly classified by the MLP classifier without attribute subset selection is done. Performing attribute subset selection to develop the model using MLP classifier decreases its performance in cholera dataset and improves its performance in malaria dataset. However, the time required to build the model by applying attribute subset selection is faster. For example, the times needed to build the models using all attributes are 1.19, 3.16 and 4.36 seconds for cholera, malaria and TB case respectively. But, this time is reduced to 0.9, 1.35 and 2.08 seconds for cholera, malaria and TB case respectively, after attribute subset selection is done on the data.

4.5. Empirical Analysis of Experiments

To develop the intended predictive models, ID3, J48, SMO and MLP algorithms are applied on the sample datasets. In the given sample datasets, a total of 24 models are developed by using all and best filtered attributes. Gain ratio attributes evaluator with ranker search method is used to filter out best and top ranked attributes for model development in each sample dataset. In each sample dataset four different experiments are conducted using the four classification algorithms, which are chosen by the researcher for the conducted experimental research work. All the models developed by selecting algorithms are tested using 10-fold cross validation evaluation method. The experiments are designed to investigate which algorithm is highly accurate on developing predictive models in the three datasets. The investigation also checks influences of instance numbers and attributes variations on the performance of algorithms and evaluates the effect of attribute subset selection on the overall performance of the algorithms. From all experiments we showed that, prediction accuracy of all algorithms applied on each sample dataset is somewhat different when the size of the data, number of attributes and types of the data are different. In this investigation change on performances of algorithms are shown by applying attribute subset selection and using a single algorithm classification for developing models in three separate datasets with different attribute numbers and number of instances.

4.5.1. The Effect of Attribute Subset Selection

Attribute subset selection does not improve the performance of ID3 in cholera dataset, improves its performance on malaria dataset and has not any change on TB dataset whether attribute subset selection is done or not. Table4.12 shows the accuracy of ID3 in the three datasets before and after attribute subset selection is done.

Table4. 12:ID3 accuracy changes due to attribute subset selection

Dataset Name	Accuracy of ID3 with all Attributes	Accuracy of ID3 with filtered Attributes	Changes
Cholera dataset	81.14%	78.48%	-2.66%
Malaria dataset	98%	98.45%	+0.45%
TB dataset	99.2%	99.2%	0%

Due to attribute subset selection, the sensitivity (TP) rate, precision, recall, F-measure, correctly classified instances and accuracy of ID3 model is decreased, whereas specificity (FP) rate and ROC area of the model is increased compared with the results of the experiments conducted using all attributes in the cholera dataset. In the same experiment when the sample dataset is changed from cholera to malaria its performance is also increased when attribute subset selection is applied to the data. However, ID3s performance is unchanged on TB dataset whether all or best selected attributes are applied on the classifier. Figure4.2 shows graphical representation of ID3 classifier models before and after attribute subset selection is done in the three sample datasets.

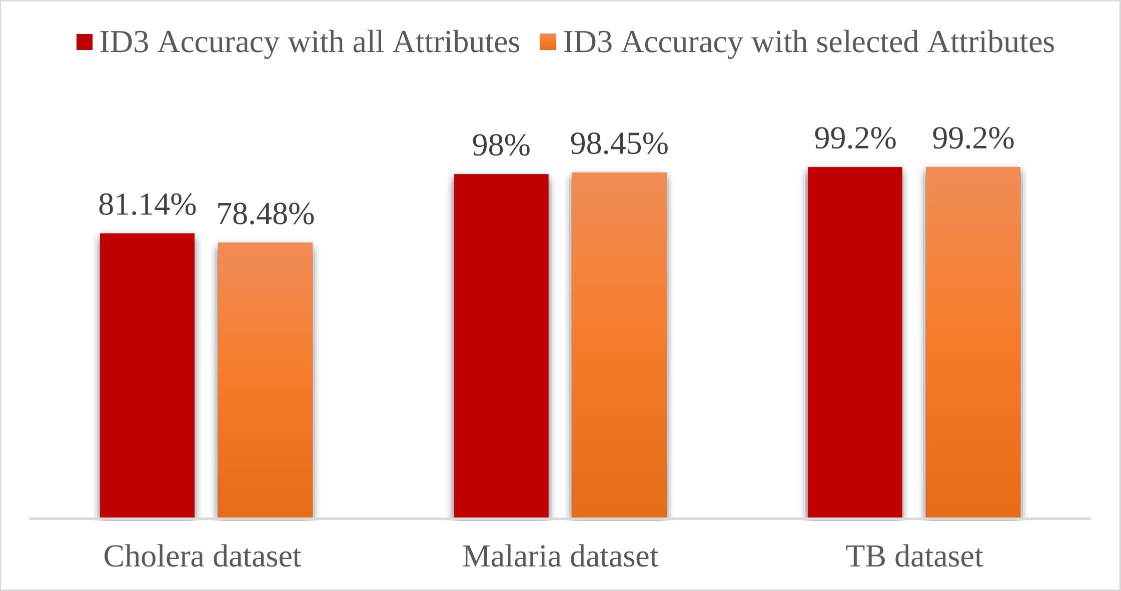


Figure4. 2: Performance of ID3 classifier before and after attribute subset selection

Similarly, the performance of J48 classifier is reduced when attribute subset selection is applied to the cholera dataset. From the experiment results in a cholera sample dataset attribute subset selection decreases true positive (TP) rate, precision, recall, F-measure, the time required to develop the model and prediction accuracy, but increases false prediction (FP) rate and ROC area. However, performance of J48 classifiers has similar accuracy on malaria and TB datasets before or after attribute subset selection is done on the classifier. Table 4.13 showed accuracy variations of a J48 classifier before and after attribute subset selection is done using gain ratio attribute evaluator in the three sample datasets.

Table4. 13:J48 classifier accuracy variation on attribute subset selection

Dataset Name	J48 Accuracy with all Attributes	J48 Accuracy with selected Attributes	Changes
Cholera dataset	82.66%	81.14%	-1.52%
Malaria dataset	98.55%	98.55%	0%
TB dataset	99.2%	99.2%	0%

The performance of SVM with SMO classifier has not shown any improvement, whether attributes subset selection is done on the three datasets. Rather, its accuracy, Precision, recall, F-measure, true positive (TP) rate and ROC area are decreased on malaria dataset and remain the same in the other two datasets. But its false positive (FP) rate is increased in malaria dataset. The predictions of ANN using MLP classifier on the three datasets show improvements on malaria dataset after attribute subset selection is done and reduces its performance on cholera and remain the same on TB dataset. Table4.14 and figure4.3 shows variations on MLP classifier in the three sample datasets before and after attribute subset selection is applied.

Table4. 14: MLP classifier accuracy variation on attribute subset selection

Dataset Name	MLP Accuracy with all Attributes	MLP Accuracy with selected Attributes	Changes
Cholera dataset	80.19%	79.62%	-0.57%
Malaria dataset	98.45%	98.55%	+0.1%
TB dataset	99.2%	99.2%	0%

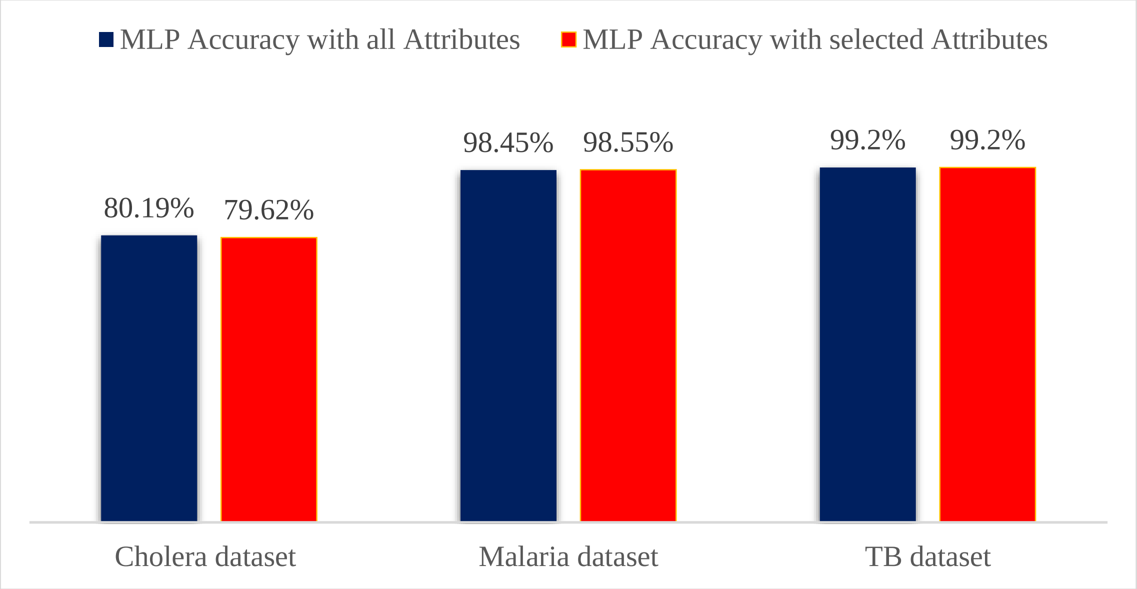


Figure4. 3: Performance of MLP classifier before and after attribute subset selection

4.5.2. The Effect of Dataset Variation on Classifiers Performance

Almost all data mining classifiers show different prediction accuracy in different types of datasets, features of datasets and size of datasets. From the 24 experiments conducted in the study, the researcher has shown different prediction accuracies using four separate classification algorithms in the three different sample datasets. For example, an accuracy of ID3 in cholera dataset is 81.14% by applying all attributes on the classifier, but its accuracy is increased to 98% in a malaria dataset with all attributes applied on the classifier using the default parameters for both experiments. On the other hand, the performance of J48 classifier is 82.66% in cholera dataset and 98.55% in malaria dataset. These variation shows prediction accuracy of a classifier is changed, when the dataset type is different, data attributes are removed from the list and classifier parameters are modified.

Prediction accuracy of all classifiers used for model development is also changed when the test options of the classifiers are different. The difference is shown in the table below in which experiments are done using 10-fold cross validation and percentage split test options (splitting data by 70:30 ratios, means 70% of the data is used for training and the remaining 30% is used for testing). From experimental results, we showed that the performances of classifiers are improved if the test option is a percentage split rather than 10-fold cross validation. Table 4.15 and figure 4.4 shows prediction accuracy variations on two test options types such as 10-fold cross validation and percentage split (70: 30 ratio) test options, which could be applied in malaria dataset.

Table4. 15: Performance of classifiers in 10-fold cross validation and percentage split

Classifier Name	Accuracy of the four classifiers with 10-fold cross validation	Accuracy of the four classifiers with percentage split (70:30 ratio)
ID3	98%	98.18%
J48	98.45%	99.39%
SMO/SVM	98.55%	99.39%
MLP	98.45%	99.39%

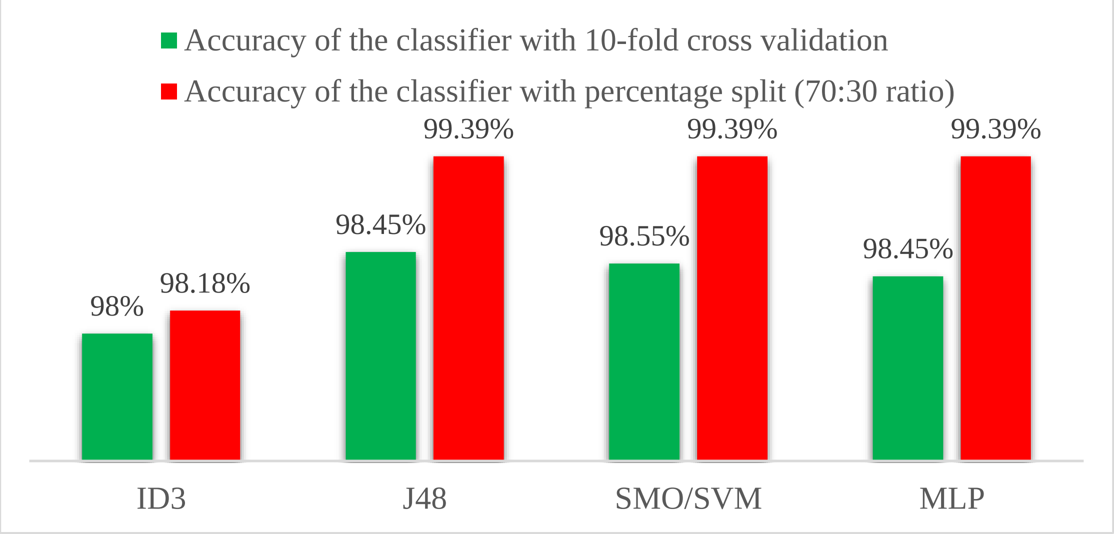


Figure4. 4: Accuracy of classifiers with 10-fold cross validation and percentage split

4.6. Model Comparison

The main purpose of the study is identifying the prediction accuracy of the four predominantly used classification algorithms such as ID3, J48, SVM and MLP Feed-Forward Neural Network using three different healthcare sample datasets i.e. cholera, malaria and TB, compare the individual performances of each algorithm in the three sample datasets and select the best algorithm for each dataset to develop prediction models. Based on this the prediction accuracy of the four algorithms is checked.

The efficiency or performances of most classifiers or algorithms are highly related to the number of attributes found in the dataset and size of the dataset itself. When attribute subset selection is done or number of attributes are reduced from the dataset, that attribute becomes high valuable for developing the required classification or predictive model. However, when the number of attribute are maximum only some attributes are highly important for developing the model and the rest are only used to increase the size and dimensionality of the data rather than increasing its efficiency. In addition to this number of instances participated in model development has also high relevance for the efficiency of an algorithm. Therefore, efficiency of an algorithm has highly maximum with limited number of number of attributes and maximum number of sample instances in the dataset than large number of attributes and limited number of sample instances participated for developing the model.

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. In each experiment confusion matrix was applied to check how much data was correctly classified in true positive (TP) correctly classified as positive cases and true negative (TN) correctly in the negative diagnosis case and also false positive (FP) healthy person falsely classified in positive case and false negative (FN) patient falsely classified as healthy individual in the study. The Confusion matrix for two possible outcomes (positive) and (negative) in each dataset of the study is described in the following three separate table with an experiment done using J48 classifier.

Table4. 16: Confusion matrix for Cholera dataset

	Predicted Positive	Predicted Negative
Actual Positive	414 (TP)	24(FN)
Actual Negative	67 (FP)	20 (TN)

Table4. 17: Confusion matrix for malaria dataset

	Predicted Positive	Predicted Negative
Actual Positive	761(TP)	2(FN)
Actual Negative	14 (FP)	323(TN)

Table4. 18: Confusion matrix for TB dataset

	Predicted Positive	Predicted Negative
Actual Positive	1017(TP)	3(FN)
Actual Negative	9 (FP)	363(TN)

To calculate TP (sensitivity), TN (specificity), FP, FN Precision, recall, F-measure, ROC area and classification accuracy the confusion matrix results are highly applied. For instance, the following are the general formulas to calculate them.

Classification accuracy= (TP + TN) / (TP + TN + FP + FN)

Error rate= (FP + FN) / (TP + TN + FP + FN)

Precision: (or Positive predictive value) proportion of predicted positives which are actual positive

Precision =TP / (TP + FP)

Recall: proportion of actual positives which are predicted positive

Recall =TP / (TP + FN)

Sensitivity: (True positive rate) proportion of actual positives which are predicted positive

Sensitivity =TP / (TP + FN)

Specificity (True negative rate) proportion of actual negative which are predicted negative

Specificity =TN / (TN + FP)

F-measure: harmonic mean between precision and recall

F-measure= 2 (precision * recall) / (precision + recall)

ROC curve: receiver operating characteristic curve: 2-D curve parameterized by one parameter of the classification algorithm, e.g. some threshold in the ≪ true positive rate / false positive rate ≫ space

Due to this the prediction ability of ID3, J48, SVM/SMO and ANN/MLP are different in different sample datasets. For example, the prediction ability of ID3 is 81.14%, 98% and 99.2% in cholera, malaria and TB patient data sets respectively. The same is true for the other three algorithms such as J48, SVM and ANN/MLP. Table 4.16 briefly shows the performance of each algorithm in the three datasets. To select an algorithm for predictive model development algorithms prediction accuracy must be considered in applying the model.

Table4. 19: Performance of all the four algorithms in the three datasets

Dataset Name	ID3 classifier accuracy	J48 Classifier accuracy	SVM Classifier accuracy	MLP Classifier accuracy
Cholera patient dataset	81.14%	82.66%	83.43%	80.19%
Malaria patient dataset	98%	98.55%	97.27%	98.45%
Tuberculosis (TB) patient dataset	99.2%	99.2%	99.2%	99.2%

As we have seen from the table the classification accuracy of the four classifiers are different in each dataset and each dataset has different in their feature structure, number of instances, types of variables, but they are tested using the same parameters of the classifier for all datasets. Graphically, this difference has been shown in the following bar graph that indicates which classifier best performs in which dataset and helps to select the best model from the developed models in the previous experiments.

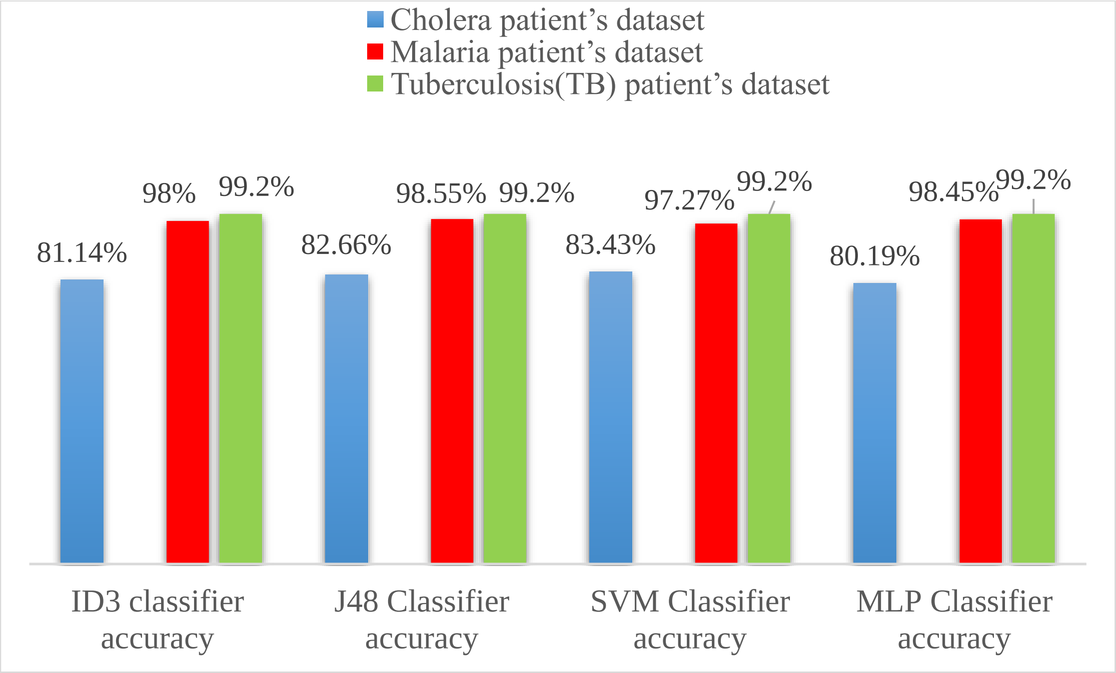


Figure4. 5: Classification accuracy of the four classifiers in the three datasets

From the table and the graph above an accuracy of all the four classifiers are between 80% and 85% for cholera dataset, their performance is increased to 98% up to 99.2% in malaria and Tb patient data sets.

Performance of SVM is 83.43%, which is higher classification accuracy in cholera dataset. Therefore, a model, which is developed using SVM, has been selected as a good model for the classification of cholera patient data set. The next best model for cholera dataset has J48 classifier with an accuracy of 82.66%, and then ID3 with an accuracy of 81.14% and the least predictive model for cholera patient dataset has MLP with an accuracy of 80.19%.

On the other hand, the best predictive model, which is selected for the classification of malaria dataset, is J48 classifier with an accuracy of 98.55%, then the next best model in this dataset has MLP with an accuracy of 98.45% and the least accuracy predictive model developed in the dataset compared with other models is SVM classifier with an accuracy of 97.27%. Lastly, for the classification of TB dataset almost all models have the same performance with an accuracy of 99.2%, but the time taken to build the model is different in each algorithm. For example, the time required for model development using ID3, J48, SVM and MLP are 0, 0, 0.17 and 4.36 seconds respectively by applying all attributes in the conducted experiments. Due to this, the researcher is recommended J48 classifier, because J48 classifier has higher prediction accuracy in other sample datasets compared with ID3 and shows the prediction results in a hierarchical tree structure. J48 classifier is a highly familiar classification algorithm compared to other classifiers applied in the conducted research.

4.7. Generated Rules from Experiments

In the conducted research, rules are generated by using ID3 and J48 decision tree classifiers based on their prediction accuracy. Therefore, the model developed using ID3 with all attributes are the best model for predicting occurrences of cholera on patients and the models developed using J48 classifier with all attributes are the best models for predicting occurrences of malaria and TB on patients. Based on this assumption a total of twenty-seven rules are selected. Seven rules are selected from cholera dataset generated by ID3 classifier and twenty separate rules are selected from malaria and TB datasets by J48 classifier ten rules for each dataset. All selected rules are acceptable by domain experts. The rules are listed and discussed below.

4.7.1. Rules of Cholera Dataset

In this dataset ID3 classifier generates a total of 23 rules. From this rule seven rules are selected to develop the model for predicting occurrences of cholera on patients.

Rule1. IF Dehydration Status=No THEN Cholera Status=Negative (81/47)

The above rule gives correct results for 47 (58.02%) of the 81 cases it covers. It shows patients who do not have any dehydration status are likely to be Cholera Negative.

Rule2. IF Watery Diarrhea=Yes AND Dehydration Status=No THEN Cholera Status=Negative (80/47)

The second rule gives correct result for 47 (58.75%) out of 80 cases. Patients not having dehydration status problems, but only have watery diarrhea are probably Cholera Negative.

Rule3. IF Weight Loss=Yes AND Dehydration Status=Severe AND Watery Diarrhea=Yes THEN Cholera Status=Positive (97/79)

The rule stated that a patient having weight loss, watery diarrhea and has severe dehydration status problems are most likely Cholera positive. It gives the correct result for 79 (81.44%) of the 97 cases tested by ID3 classifier.

Rule4. IF Watery Diarrhea=Yes AND Appetite Loss=No AND Dehydration Status=No THEN Cholera Status=Negative (75/44)

The fourth rule stated that a patient having watery diarrhea and has not any loss of appetite and dehydration status issues the disease likely not to be cholera. It gives correct results for 44 (58.67%) of 75 cases in the experiment.

Rule5. IF Watery Diarrhea=Yes AND Dehydration Status=Some AND Diarrhea per day= [7-11] AND Weight Loss=No THEN Cholera Status=Positive (99/91)

The fifth rule stated that a patient having watery diarrhea, severe dehydration status and has a diarrhea per day between 7 and 11 but does not have any loss of weight issues are most likely be cholera positive. It gives correct results for 91 (91.92%) of 99 cases in the experiment.

Rule6. IF Watery Diarrhea=Yes AND Weight Loss=No AND Appetite Loss=No AND Dehydration Status=No THEN Cholera Status=Negative (72/42)

The sixth rule gives correct result for 42 (58.33%) out of 72 cases. Patients having watery diarrhea, but not any weight loss, appetite loss and dehydration status problems are probably Cholera Negative patients.

Rule7. IF Diarrhea per day= [1-5] AND Appetite Loss=No AND Dehydration Status=No THEN Cholera Status=Negative (67/39)

The last and seventh rule gives correct result for 39 (58.33%) of 67 cases covered. It shows patients having diarrhea per day between 1 and 5 and have not any appetite loss and dehydration status problems are probably Cholera Negative patients.

4.7.2. Rules of Malaria Dataset

In this sample dataset of the study a total of 47 rules is generated by J48 classifier. From this ten best rules are selected by the researcher to develop the predictive model for predicting occurrences of malaria in a patient. All the selected rules are getting acceptance from domain experts.

Rule1: IF Fever=Yes AND Joint pain=Yes AND Tiredness=Yes THEN Malaria Status=Positive (773/761)

The first rule in malaria dataset stated that a patient having Fever, Joint pain and Tiredness problems are probably Malaria Positive. It gives the correct result for 761 (98.45%) of 773 cases covered.

Rule2: IF Fever=Yes AND Headache=Yes AND Joint pain=Yes AND Tiredness=Yes THEN Malaria Status=Positive (773/761)

The second rule gives the correct result for 761 (98.45%) out of 773 cases tested by the classifier. It stated that a patient having Fever, Headache, Joint pain and Tiredness are probably Malaria Positive patient.

Rule3: IF Fever=Yes AND Chill=Yes AND Joint pain=Yes AND Tiredness=Yes THEN Malaria Status=Positive (773/761)

The third rule stated that a patient having Fever, Chill, Joint pain and Tiredness are most likely Malaria Positive. It gives the correct result for 761 (98.45%) of 773 cases tested in the experiment.

Rule4: IF Fever=Yes AND Headache=Yes AND Chill=Yes AND Joint pain=Yes AND Tiredness=Yes THEN Malaria Status=Positive (773/761)

The fourth rule gives correct result for 761 (98.45%) of 773 sample instances tested by the classifier. This stated that a patient having Fever, Headache, Chill, Joint pain and Tiredness are highly Malaria Positive patients.

Rule5: IF Fever=Yes AND Joint pain=Yes AND Tiredness=Yes AND Coma=No THEN Malaria Status=Positive (769/757)
The fifth rule stated that a patient having Fever, Joint pain and Tiredness but not be in a coma there probably Malaria Positive. It gives the correct result of 757 (98.44%) of 769 instances tested by the classifier.

Rule6: IF Fever=Yes AND Headache=Yes AND Joint pain=Yes AND Tiredness=Yes AND Coma=No THEN Malaria Status=Positive (769/757)
The sixth rule stated that a patient having Fever, Headache, Joint pain and Tiredness but not be in a coma most probably Malaria Positive. It gives the correct result of 757 (98.44%) of 769 instances tested by the classifier.

Rule7: IF Fever=Yes AND Chill=Yes AND Joint pain=Yes AND Tiredness=Yes AND Coma=No THEN Malaria Status=Positive (769/757)
This rule gives the correct result of 757 (98.44%) of 769 instances tested by the classifier. The seventh rule stated that a patient having Fever, Chill, Joint pain and Tiredness but not be in a coma most likely Malaria Positive.

Rule8: IF Fever=Yes AND Joint pain=Yes AND Tiredness=Yes AND Sweat=Yes AND Severe Anemia=Yes AND Shock=No AND Renal Failure=No THEN Malaria Status=Positive (767/755)
The eighth rule stated that a patient having Fever, Joint pain, Tiredness, Sweat and Severe Anemia but not have any shock and renal failure problems are high probably of Malaria Positive. It correctly predicted 755 (98.2%) of 769 cases tested by the classifier.

Rule9: IF Fever=Yes AND Headache=Yes AND Joint pain=Yes AND Tiredness=Yes AND Sweat=Yes AND Severe Anemia=Yes AND Renal Failure=Yes AND Shock=No AND Coma=No AND Confusion=No THEN Malaria Status=Positive (767/755)
The ninth rule stated that a patient having Fever, Headache, Joint pain, Tiredness, Sweat, Severe Anemia and Renal failure, but not has any shock, Coma and Confusion problems

are high probability of Malaria Positive. It correctly predicted 755 (98.2%) of 769 cases tested by the classifier.

Rule10: IF Fever=Yes AND Chill=Yes AND Joint pain=Yes AND Tiredness=Yes AND Sweat=Yes AND Severe Anemia=Yes AND Renal Failure=Yes AND Shock=No AND Coma=No AND Confusion=No AND Vomiting=No THEN Malaria Status=Positive (767/755)

The tenth rule stated that a patient having Fever, Headache, Joint pain, Tiredness, Sweat, Severe Anemia and Renal failure, but not have any shock, Coma Confusion and vomiting problems are high probability of Malaria Positive. It correctly predicted 755 (98.2%) of 769 cases tested by the classifier.

4.7.3. Rules of TB Dataset

Like other sample datasets experimented in the study a total of 63 rules are generated by J48 classifier. From this only ten best rules are selected to develop the model for predicting occurrences of TB in a patient. All the selected rules are accepted by domain experts.

Rule1. IF Chest Pain=Yes AND Productive Sputum=Yes THEN TB Status=Positive (906/899)

The first rule of TB dataset stated that a patient having Chest Pain and Productive Sputum (Bloody Sputum) has a high probability of TB positive. It gives the correct result of 899 (99.23%) of 906 sample instances tested by the classifier to develop the model.

Rule2. IF Chest Pain=No AND Productive Sputum=No AND Night Sweating=Yes AND Weight Loss= No AND HIV Test Result= NR THEN TB Status=Negative (486/481)

The second rule stated that a patient has not Chest Pain, Productive Sputum (Bloody Sputum), Night Sweating, Weight Loss and non-reactive HIV test result has a high probability of TB Negative. It gives the correct result of 481 (99 %) of 486 sample instances tested by the classifier to develop the model.

Rule3. IF Night Sweating=Yes AND Productive Sputum=Yes AND Shortness of Breath =No AND Loss of Appetite =No THEN TB Status=Positive (906/900)

The above rule stated that a patient having night sweating and productive sputum but does not have shortness of breath and loss of appetite are positive in TB status. It gives correct result for 900 (99.34%) out of 906 sample instances tested by J48 classifier.

Rule4. IF Cough (>2wks) =Yes AND Productive Sputum=Yes AND Loss of Appetite =Yes AND Headache=No AND Fever=No THEN TB Status=Positive (906/899)

Rule four stated that a patient having a cough which stayed more than two weeks, productive sputum and loss of appetite but does not any headache or fever are high probability of TB positive status. This gives the correct result of 899 (99.23%) of 906 cases covered in the test.

Rule5. IF Chest Pain =Yes AND Productive Sputum=Yes AND Weight Loss=Yes AND Night Sweating=No AND Fever=No THEN TB Status=Positive (906/901)

This rule stated that a patient having chest pain, productive sputum and weight loss but does not have any sweating at night and fever are likely TB positive. This gives correct result for 901 (99.45%) of 906 cases tested by the classifier.

Rule6. IF HIV test Performed=Yes AND HIV test result=R AND Productive Sputum=Yes AND Night Sweating=Yes THEN TB Status=Positive (906/899)

The sixth rule stated that a patient having performed HIV tests and its test result is reactive as well as productive sputum and sweating at night is a high degree of TB positive. It produces correct results for 899 (99.23%) of 906 cases tested by J48 classifier.

Rule7. IF HIV test result=R AND Productive Sputum=Yes AND Cough (>2wks) =Yes AND Shortness of Breath= Yes THEN TB Status=Positive (906/901)

The above rule showed a patient having a reactive HIV test result, productive sputum, cough stayed more than two weeks and have a breathing problem are high TB positive status. This gives correct result for 901 (99.45%) of 906 cases tested by the classifier.

Rule8. IF HIV test Performed=Yes AND HIV test result=NR AND Productive Sputum=Yes AND Weight Loss=Yes AND Loss of Appetite =Yes THEN TB Status=Positive (906/900)

The eighth rule stated that a patient having performed HIV tests and the HIV test result is reactive as well as a patient has productive sputum, weight loss and loss of appetite are most likely TB status Positive. This gives correct result for 900 (99.34%) of 906 cases.

Rule9. IF HIV test Performed=Yes AND HIV test result=NR AND Productive Sputum=Yes AND Weight Loss=Yes AND Night Sweating=Yes THEN TB Status=Positive (906/900)

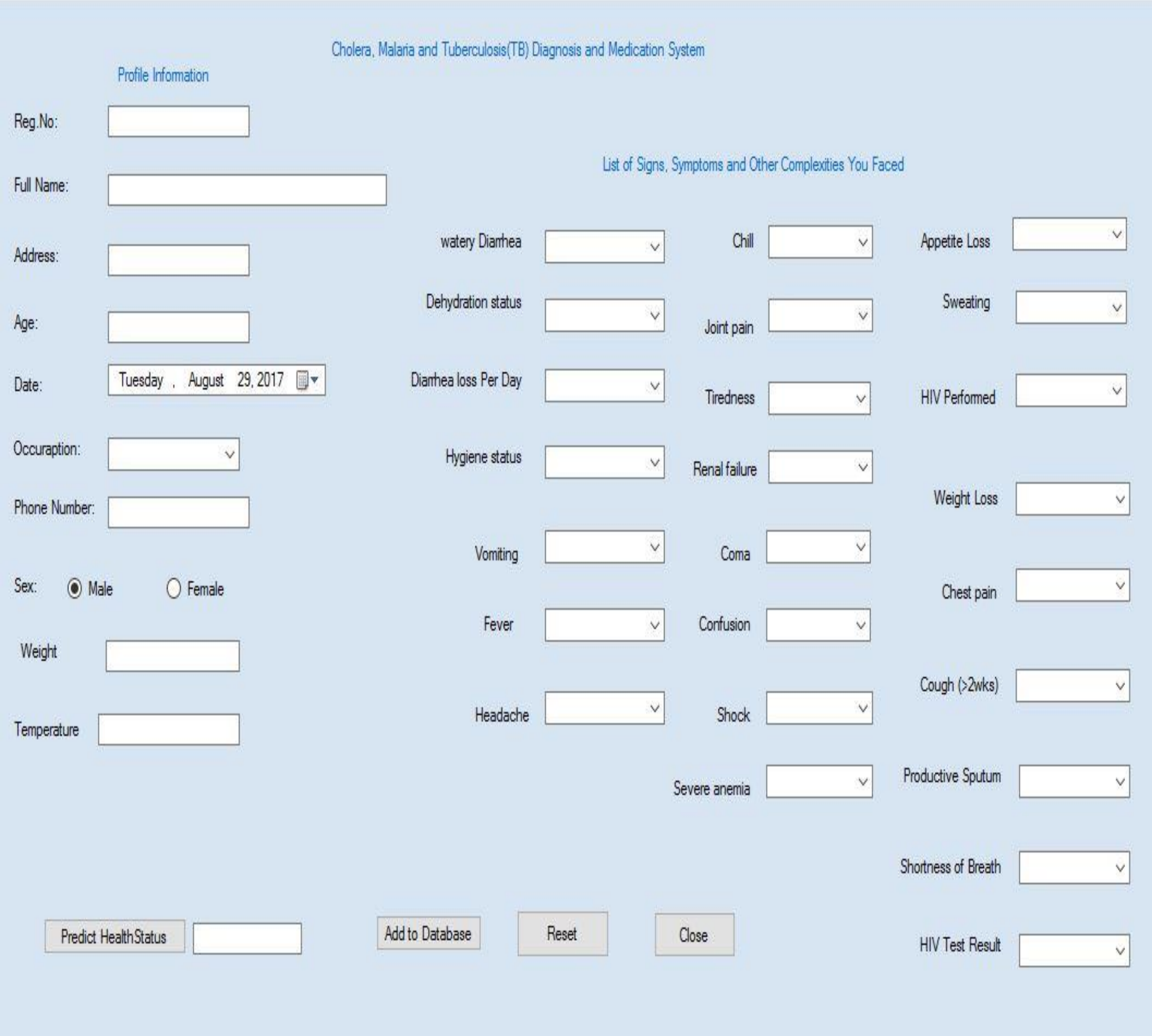
The ninth rule stated that a patient having performed HIV tests and the HIV test result is reactive as well as a patient has productive sputum, weight loss and sweating at night are high probability of TB Positive. This gives correct result for 900 (99.34%) of 906 cases.

Rule10. IF Cough (>2wks) =Yes AND HIV test result=R AND Productive Sputum=Yes AND Weight Loss=Yes AND Night Sweating=Yes THEN TB Status=Positive (906/901)

The last rule of this dataset stated that a patient having a cough that stayed more than two weeks, has reactive HIV test result, productive sputum, weight loss and has sweating at night are more likely TB positive. This gives correct result for 901 (99.45%) of 906 cases checked by the classifier.

4.8. Graphical User Interface Development

For the development of graphical user interface Microsoft visual studio 2013 is used. This prototype graphical user interface is developed based on the best models chosen above for classification of the three datasets using all attributes. Before developing the graphical interface, the researcher is generated rules by using one of the classifier called J48 algorithm. The rules used by the researcher are important to design the graphical user interface for predicting occurrences of cholera, malaria and TB disease in patients. Figure 4.2 showed the graphical user interface developed for predicting occurrences of cholera, malaria and TB disease in patients.


Date: Tuesday , August 29, 2017

Sex: ● Male ○ Female

Figure4. 6: Graphical User Interface for the Conducted Research

CHAPTER FIVE

5. Conclusion and Recommendation

5.1. Conclusion

Infectious diseases are the leading cause of illness and death in low and middle-income countries. In Ethiopia, this disease covers the top 10 leading cause of inpatient as well as outpatient visits in different parts of the country. Due to this, an increasing amount of medical data is recorded every day from the healthcare sector. Knowledge extraction on those data wealthy, but knowledge poor sector, using data mining methods are capable of providing decision support for the healthcare practitioners and to discover relevant patterns from patient diagnostic activities, prescriptions, treatments or clinic information management data.

The study analyzed infectious diseases patient datasets and developed a predictive model for effective diagnosis and treatment activities. In this investigation, a total of 3017 sample instances were taken from two government hospitals for experimentation by considering three major infectious disease types covered in the investigation of the study. Those include 525 instances of cholera, 1100 instances of malaria and 1392 instances of TB cases. The experiment is done using three popular data mining, classification algorithms such as decision Tree, SVM and ANN.

The performances of the models are evaluated using accuracy, precision, recall, F-measure, true positive (TP) and false positive (FP) rate and ROC area. 10-Fold Cross Validation is adopted as a test option for random sampling of the training and test data samples. In the four experiments, each experiment contains six separate models that are performing well in predicting cholera, malaria as well as TB disease patients. The most effective model of predicting patients with cholera disease appears to be an SVM classifier implemented on all attributes with a classification accuracy of 83.43%, J48 classifier for malaria and TB diseases implemented on all attributes with a classification accuracy of 98.55% and 99.2% respectively.

The study showed that data mining techniques are highly important to develop an effective predictive model for infectious diseases datasets. The outcome of the study can be used as an assistant tool for physicians to support them and to make more consistent diagnoses. Furthermore, this is also important to identify patients of infectious diseases by considering attributes that increase the probability of a particular infectious disease in the patient's body.

5.2. Recommendation

In this research, a number of tasks are completed to find the possible applicability of data mining technology on predicting occurrences of cholera, malaria and TB in patients. Even though the study, which is conducted has dedicated to the academic exercise, its results are hopeful to be applied in addressing practical problems on prevention and control of infectious disease. The researcher forwarded the following recommendations based on the result of this study:

- ✓ The study is trying to develop predictive models and graphical interface for the correct occurrence prediction of cholera, malaria and TB disease. However, domain experts desire the development of knowledge based system (KBS) for the future prediction of the disease. This should be a future research direction.

- ✓ The researcher used only single algorithm for the development a particular model. But, in the future integrated or multiple algorithm integration techniques will provide better result than applying single algorithms for model development.

- ✓ The samples do not represented all infectious disease patients of the country, because it is collected only from two government hospitals located in the Bahir dear city. However, experiments on data mining require very huge representative sample datasets from different regions of the country and involve private hospitals for collecting samples to get highly effective results which makes the diagnosis and treatment activities are time saver and accurate.

- ✓ Hospitals have record medical information with manual recording mechanisms, which is time consuming and passes unsecure information management practices. Therefore, electronic medical record management system for secure information management, effective diagnosis, prognosis and treatment is mandatory.

References

Abdur , R. (2011). A questionnaire survey on infectious disease among hospital patients in Kushtia and Jhenaidah, Bangladesh. International Journal of Genetics and Molecular Biology Vol. 3(9), , pp. 120-134.

Abel , D. (2011). Designing A Predictive Model for Heart Disease Detection Using Data Mining Techniques. Masters thesis Addis Ababa University .

Abirami et al. (2013). A Study on Analysis of Various Data mining Classification Techniques on Healthcare Data. International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 7 , PP 604-607.

Alessio et al. (2015). Chapter2 Health Trends of Communicable Diseases. Geneva: Springer International Publishing.

Amato et al. (2013). Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine , pp 48-58.

Asia , N. (2012). Mining Patients' Data for Effective Tuberculosis Diagnosis: The Case of Menelik II Hospital. Masters Thesis, Addis Ababa University .

Azman al et. (2017). The incubation period of cholera: a systematic review. Journal of Infection .

Brownlie, J., Peckham, C. C., Waage, J., Woolhouse , M. O., Lyall, C., Meagher, L., et al. (2006). Foresight. Infectious Diseases: preparing for the future Future Threats. London: Office of Science and Innovation.

Charles , P. (2017). Retrieved from infectious disease center / infectious disease a-z list / tuberculosis (tb) facts index / tuberculosis (tb) facts article:

Colin , S. (2000). The CRISP-DM Model:The New Blueprint for Data Mining. Journal of data warehousing Volume 5 Number 4 , PP 13-22.

DACAE. (2010). Standard Treatment Guideline for General Hospitals. Addis Ababa, Ethiopia: Drug Administration and Control Authority of Ethiopia .

David , L., & Dursun , D. (2008). Advanced Data Mining Techniques. Berlin Heidelberg: Springer.

Desikan, P., Hsu, K.-W., & Srivastava, J. (2011). Data Mining for Healthcare Management. 2011 SIAM International Conference on Data Mining. Arizona, USA: Hilton Phoenix East.

Divya , T., & Sonali , A. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology Vol.5, No.5 , pp. 241-266.

Divya, T., & Sonali , A. (2014). A survey on pre-processing and post-processing techniques in data mining. International Journal of Database Theory and Application Vo.7 No.4 , pp 99-128.

Durairaj, M., & Ranjani, V. (2013). Data Mining Applications In Healthcare Sector: A Study. International Journal of Scientific & Technology Research Volume 2, Issue 10 , PP.29-35.

Fayyad et al. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. COMMUNICATIONS OF THE ACM Vol. 39, No. 11 , PP 27-34.

FMoH. (2010). Health and Health Related Indicators: 2008/9. Addis Ababa: Federal Ministry of Health:Ethiopia.

FMoH. (2010). Communicable Diseases: Part 1 General principles, vaccine-preventable diseases and malaria, Blended Learning Module for the Health Extension Programme. Addis Ababa: Federal Ministry of Health:Ethiopia.

FMoH. (2010). Communicable Diseases: Part 2 Tuberculosis and leprosy, Blended Learning Module for the Health Extension Programme. Addis Ababa: Federal Ministry of Health: Ethiopia.

FMoH. (2010). Communicable Diseases: Part 4 Other Diseases of Public Health Importance and Surveillance, Blended Learning Module for the Health Extension Programme. Addis Ababa, Ethiopia: Federal Ministry of Health: Ethiopia.

FMoH. (2011). First Ethiopian National Population Based Tuberculosis Prevalence Survey. Addis Ababa. Ethiopia: Federal Ministry of Health: Ethiopia.

Gaurav, T., & Ashwini, S. (2014). Study of classifiers in data mining. International Journal of Computer Science and Mobile Computing Vol.3 Issue.9 , pg. 263-269.

GIDEON. (2010). Infectious Diseases of Haiti - 2010 edition . California, USA: GIDEON Informatics, Inc,.

Guyon, I., & Andre , E. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 , pp 1157-1182.

Hailu, T. G. (2015). Comparing Data Mining Techniques in HIV Testing Prediction. Intelligent Information Management , pp 153-180.

Harshank, G., & Pushpendra, K. P. (2014). A framework for Data Analysis using Data Mining. IOSR Journal of Engineering , Vol. 04, Issue 03 , PP 33-35.

Jiawei , H., & Micheline, K. (2006). Data Mining Techniques and Concepts Second Edition. USA: Morgan Kaufmann Publishers.

Jiawei al et. (2012). Data Mining Concepts and Techniques,3rd Edition. Waltham, USA: Morgan Kaufmann Publisher.

Joyce , J. (2002). Data mining: a conceptual overview. Communications of the Association for Information Systems Volume 8 , 267-296.

Kamran, E. S. (2013). Importance of Data Preprocessing For Improving Classification Performance on CAC Data Set. Graduate School of Science and Engineering Kadir Has University .

Kariuki et al . (2016). Towards Effective Data Preprocessing for Classification Using WEKA. International Journal of Science and Research, Volume 5, Issue 8 , pp 210-214.

Kotsiantis et al. (2006). Data Preprocessing for Supervised Leaning. International Journal of Computer Science Volume 1 Number 1 , pp 111-117.

Kumar, L., & Padmapriya, A. M. (2012). ID3 Algorithm Performance of Diagnosis For Common Disease. International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 5 , pp 57-62.

Linda et al. (2003). Infection Prevention Guidelines for Healthcare Facilities with Limited Resources. Maryland, USA: JHPIEGO Corporation.

Lozano R, et al. (2013). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010. a systematic analysis for the global burden of disease study , PP 2095–2128.

Mariammal et al. (2014). Major Disease Diagnosis and Treatment Suggestion System Using Data Maining Techniques. International Journal of Advanced Research in Computer Science & Technology, Vol. 2 Issue 1 , pp 338-341.

Mary , K. (2004). Application of Data Mining Techniques to Healthcare Data. Infection Control And Hospital Epidemiology, Statistics For Hospital Epidemiology , Vol. 25 (No. 8), pp. 690-695.

Mary , M., & Sandra , Y. (2006). Controlling Infectious Diseases. Population Bulletin

Vol. 61, No. 2 , pp 1-20.

Mary, K., & Obenshain, M. (2004). Application of Data Mining Techniques to Healthcare Data. Chicago Journals The Society for Healthcare Epidemiology of America , pp. 690-695.

Masumeh, m. A., & Peyman, B. (2016). Presenting a Hybrid Model for Early Diagnosis of Hepatitis by Applying Data Mining Techniques. International Journal of Computer & Information Technologies (IJOCIT) , pp 1-10.

Mehdi , T., Farshad , F., Mahsa , S. A., Amir , H. -M., Parisa, A. A., Ehsan , R.-D., et al. (2016). Detecting Diseases in Medical Prescriptions Using Data Mining Tools and Combining Techniques. Iranian Journal of Pharmaceutical Research vol. 15 , pp 113-123.

Michael, G., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. SIGKDD Explorations, Volume 1, Issue 1 , PP 20-33.

Mohammad et al. (2012). A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining. International Journal of Applied Engineering Research, Vol.7 No.11 .

Oguntimilehin A et al. (2015). A Review of Predictive Models on Diagnosis and Treatment of Malaria Fever. International Journal of Computer Science and Mobile Computing, Vol.4 Issue.5 , pg. 1087-1093.

Padma , P. (2004). E-Intelligence form design and Data Preprocessing in Health Care. A Master thesis Waterloo University, Ontario, Canada .

Parvez, A., Saqib, Q., & Syed, Q. A. (2015). Techniques of Data Mining In Healthcare: A Review. International Journal of Computer Applications , Volume 120 (No.15).

Pradhan. (2016). Analysis of Data Mining Techniques for Building Health Care Information System. International Journal of Engineering Technology, Management and Applied Sciences Volume 4, Issue 1 , PP 49-56.

Pradhan. (2014). Data Mining and Health Care: Techniques of Application. ISOI Journal of Engineering and Computer science Volume 1 Issue 1 , PP. 18-26.

Pradhan, M. (2015). Data Mining Techniques in Health Care Application: A Review of Survey. 2nd International Conference on Recent Innovations on Science, Engineering and Management , PP 42-57.

Richard , A. (2017). Chapter 24 Cholera, Vibrio cholerae O1 and O139, and Other

Pathogenic Vibrios. Retrieved from Medical Microbiology. 4th edition.

Ruben, D. C. (2009). Data mining in healthcare: current applications and issues. Masters Thesis in Carnegie Mellon University Australia .

Sakshi, & Sunil , K. (2015). A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining. International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 8 , PP. 5142 - 5147.

Saulat , J. (2016). Cholera – Epidemiology, Prevention and Control. Significance, Prevention and Control of Food Related Diseases , pp 145-157.

Shelly et al. (2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian Journal of Computer Science and Engineering, Vol. 2 No. 2 , PP 188-195.

Sujatha et al. (2016). A Survey of Health Care Prediction Using Data Mining. International Journal of Innovative Research in Science,Engineering and Technology Vol. 5, Issue 8, PP 14538-14543.

Sushmita , M., & Tinku, A. (2003). Data Mining Multimedia, Soft Computing,and Bioinformatics. canada: John Wiley & Sons, Inc.

ȚĂRANU, I. (2015). Data mining in healthcare: decision making and precision. Database Systems Journal vol. VI, no. 4 , PP 33.40.

Tasha et al. (2012). Data Mining Techniques in the Diagnosis of Tuberculosis. Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis Dr. Pere-Joan Cardona (Ed.), ISBN: 978-953-307-938-7, InTech , PP. 333-352.

Teketel , M. (2013). Constructing A Predictive Model for Occurrence of Tuberculosis: The Case of Menelik II Hospital and St. Peters TB Specialized Hospital. Masters thesis Addis Ababa University.

Umair , S., & Haseeb , Q. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research Vol. 12 No. 1 , PP. 217-222.

UNICEF. (2008). Humanitarian Action Report: Health, Education, Equality, Protection Advance Humanity. New York: United Nations Children's Fund.

Vijeta S et al. (2015). Malaria Outbreak Prediction Model Using Machine Learning.

International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 12 , pp 4415-4419.

WHO. (2012). Global report for research on infectious diseases of poverty. Geneva: World Health Organization .

WHO. (2014). Global Tuberculosis Report. Geneva, Switzerland: World Health Organization.

WHO. (2015). Global Tuberculosis Report 20th Edition . Geneva, Switzerland: World Health Organization.

WHO. (2013). Mortality and global health estimates. Geneva: World Health Organization.

WHO. (2017). The top 10 causes of death. Fact sheet. Retrieved from Global burden of disease.: http://www.who.int/mediacentre/factsheets/fs310/en/index.

WHO. (2012). Weekly epidemiological record. Geneva: World Health Organization.

WHO. (2016). Weekly epidemiological record. Geneva: World Health Organization.

WHO. (2014). WHO Global Malaria Programme World Malaria Report 2014. Geneva, Switzerland: World Health Organization.

WHO. (2015). WHO Global Malaria Programme World Malaria Report 2015. Geneva, Switzerland: World Health Organization .

Witten. I & Frank. E (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., San Francisco, Morgan Kaufmann.

Zhang, Y., Yi, X., Qin, L., Jianshe, M., Shuai, L., Xiaodan, L., et al. (2017). Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. BioMedical Engineering online , PP. 1-15.