ADAMA SCIENCE & TECHNOLOGY UNIVERSITY

SCHOOL OF ELECTRICAL ENGINEERING & COMPUTING

DEPARTMENT OF COMPUTING

A MASTER'S THESIS

ON

AN INTEGRATION OF CASE BASED REASONING WITH DATA MINING TECHNIQUE TO ENHANCE THE EFFECTIVENESS: IN CASE OF PNEUMONIA DISEASE DIAGNOSIS

BY

MINTESINOT ABEBE

JANUARY, 2018

ADAMA, ETHIOPIA.

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTING

AN INTEGRATION OF CASE BASED REASONING WITH DATA MINING TECHNIQUE TO ENHANCE THE EFFECTIVENESS: IN CASE OF PNEUMONIA DISEASE DIAGNOSIS

BY

MINTESINOT ABEBE

THE THESIS SUBMITTED TO SCHOOL OF GRADUATE STUDIES OF ADAMA SCIENCE AND TECHNOLOGY UNVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN SOFTWARE ENGINEERING

JANUARY, 2018

Adama, Ethiopia

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTING

AN INTEGRATION OF CASE BASED REASONING WITH DATA MINING TECHNIQUE TO ENHANCE THE EFFECTIVENESS: IN CASE OF PNEUMONIA DISEASE DIAGNOSIS

BY

MINTESINOT ABEBE

SIGNATURE PAGE (Approval sheet)

_____	_____	_____
Advisor	**Signature**	**Date**
_____	_____	_____
Chair Person	**Signature**	**Date**
_____	_____	_____
External Examiner	**Signature**	**Date**
_____	_____	_____
Internal Examiner	**Signature**	**Date**
_____	_____	_____
Head of the Department	**Signature**	**Date**
_____	_____	_____
Dean of School	**Signature**	**Date**

DECLARATION

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university.

MINTESINOT ABEBE

JANUARY, 2018

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

Dr. MESFIN ABEBE

JANUARY, 2018

ACKNOWLEGEMENT

First and foremost I would like to thank God and Holy Mother who made all things possible, and granted me success in my thesis work and entire journey. Next, I would like to forward my gratitude to my supervisor, Dr. Mesfin Abebe, for his advice, supervision, and guidance from the early stage of this thesis until its successful accomplishment. Many thanks go in particular to Mr. Mohamed Wazhi and Mr. Tagel Aboneh who helped me by providing important information and resources at pre stage of the research proposal. I would also like to thank Mr. Epherem who supported me to get pneumonia data set from the Adama Hospital and Medical Collage and also Mr. Solomon who collect the data set from the Medical Record Cards. I would also like to thank Adama Hospital and Medical Collage doctors who participated in evaluating the prototype system. Finally, I would like to thank my family and friends for their prayers and words of encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF LISTINGS

ABBREVIATION AND ACRONYM

AI	Artificial Intelligence
ANN	Artificial Neural Network
ARM	Association Rule Mining
CAP	Community-Acquired Pneumonia
CBR	Case Based Reasoning
CBRDM	Case Based Reasoning Data Mining
CRISP-DM	Cross Industry Standard Process for Data Mining
DM	Data Mining
EMR	Electronic Medical Records
FL	Fuzzy Logic
GA	Genetic Algorithm
GAIA	Group for Artificial Intelligence Applications
HSQLDB	Hypersonic Structural Query Language Database
JCOLIBRI	Java Class Ontology Libraries Integration for Building Reasoning Infrastructure
JDBC	Java Database Connectivity
KBS	Knowledge Based System
KDD	Knowledge Discovery in Databases
KNN	K Nearest Neighbor
MF	Membership Function
ML	Machine Learning
MRNS	Medical Record Numbers
NB	Naïve Bayes
ODANB	One Dependency Augmented Naïve Bayes classifier

OSS	Open Source Software
OWL	Ontology Web Language
PR	Probabilistic Reasoning
RBR	Rule Based Reasoning
SPSS	Statistical Package for the Social Sciences

ABSTRACT

Nowadays nothing is more worth than health. It is the most important aspects of human life. Pneumonia is one of the health related problem which is the most killing disease in Ethiopia even in Africa. The reason for its high level of killing is the difficulty of the disease nature to treat as well as the shortage of health professionals. We examined the strengths and weaknesses of various reasoning paradigms including case-based reasoning, rule-based reasoning and data mining techniques. We discuss how to combine them to form a more robust and better-performing hybrid. In a decision support system to address the variety of tasks a user performs, a single type of knowledge and reasoning method is often not sufficient. A combination of different methods has often shown the best results. In this study Case Based Reasoning was mixed with data mining techniques and Rule Based Reasoning approaches to promote synergies and benefits beyond those achievable using CBR or other individual reasoning approaches alone.

The dataset is collected from Adama medical health science collage and hospital and different standard guideline using document review technique. The data from Medical Record Numbers are collected and analyzed to be used in Case Based Reasoning. The dataset has 1007 records and 15 fields. The industry standard CRISP-DM data mining process model used throughout this research for the purpose of preprocessed and model building. The case stored in the database, preprocessed using a data mining techniques. The dataset is further used to generate classification rule and association rule. To develop a prototype, we use mainly jcolibri frame work library for CBR prototype, Eclipse IDE, JavaFx, FXML and Navicat (MySQL database).

The evaluation of the prototype is done from two system measurement perspectives. The first is from domain expert acceptance test perspective which uses standard user acceptance criteria metrics to evaluate the prototype. We found the prototype is accepted by domain experts on average 84% which is promising result compared with other similar systems. The other standard metrics used to measure retrieval performance is recall and precision. We acquired 88% of relevant cases are retrieved out of the total assigned relevant cases by domain experts which is a promising recall result. The rules generated using data mining techniques are in support of the specific case experienced knowledge. Such a clinical decision support systems are very significant in countries like Ethiopia where shortage of health professionals are high.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

The use of computer-based decision support systems within the field of health science has over the last decades been extensively researched and tested, both in controlled environments and in clinical practice [1]. Computer-based decision support systems can be used in the field of health science in different functionalities or applications such as Assessment, Patient Monitoring, Documentation, Telemedicine, Electronic Medical Records (EMR) and Diagnosis. Diagnosis can be done by exploring the exposed symptoms, the system state, the general specification of the system and the operating environment. In diagnosis, the behavior of an observed system is checked for previously defined problem conditions to explain the current problem that the system is experiencing. Reusing previous experiences in diagnosis can result in faults being corrected more quickly and more consistently and is the method employed by CBR [2].

CBR is an emerging decision making paradigm in medical research where new cases are solved relying on previously solved similar cases [2]. Usually, a database of solved cases is provided, and every case is described through a set of attributes (inputs) and a label (output). Extracting useful information from this database can help the CBR system providing more reliable results on the yet to be solved cases. CBR is sometimes classified under Machine Learning, and supports knowledge acquisition and problem-solving. It is also sometimes associated with other technologies such as analogy, cognitive psychology modeling, machine learning, and information retrieval [3].

Case-based representations store a large set of previous cases with their solutions in the CASE BASE and use them whenever a similar new case has to be dealt with. Case based reasoning is different from other AI branches. Most expert systems stick on giving a generalized knowledge, but CBR utilizes the specific knowledge of previously experienced, concrete problem (cases) [2].

The history of CBR in AI is started in the works of Roger Schank on dynamic memory and the central role that a reminding of earlier situations and situation patterns has in problem solving and learning. The first system that might be called a case-based reasoner was the CYRUS system,

developed by Janet Kolodner, at Yale University. CYRUS was a question-answering system with knowledge base of the various travels and meetings of former US Secretary of State Cyrus Vance. Then after the CYRUS system so many CBR system were developed including MEDIATOR, PERSUADER, CHEF, JULIA, CASEY, but all these CBR systems are based on CYRUS system of case memory model [2].

Data mining techniques is one way of knowledge discovery from the database which is using different techniques such as classification, clustering, association, prediction etc. All these techniques have algorithms to perform tasks to find some hide knowledge and to use it for dedicated purpose. Data mining refers to the analysis of the large quantities of data that are stored in computers. It has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. However, systematic exploration through classical statistical methods is still the basis of data mining. Some of the tools developed by the field of statistical analysis are harnessed through automatic control (with some key human guidance) in dealing with data [4].

Pneumonia Disease is one of lower respiratory tracts contaminations and it has many sorts of causes. It is a contamination in one or both lungs. It can be caused by organisms, microbes, or infections. Pneumonia causes aggravation in our lung's air sacs, or alveoli. The alveoli load with liquid or discharge, which makes breathing difficult and points of confinement oxygen intake [5]. Impacts of pneumonia disease can range from mild to life hreatening.

Pneumonia remains the leading cause of death in children under five worldwide. It accounts for about 1.6 million deaths a year in this age group and 18% of all deaths among children under five [6], [7]. More than 99% of all pneumonia deaths occur in low- and middle-income countries [8]. South Asia and sub-Saharan Africa bear the burden of more than half of the total number of cases of suspected pneumonia among children under five worldwide. Children in low-income countries are nearly 18 times more likely to die before the age of five than children in high-income countries, due mainly to pneumonia and other acute infections [7].

In Ethiopia, pneumonia is a leading single disease killing under-five children. It is estimated that 3,370,000 children encounter pneumonia annually which contributes to 20 percent of all causes of deaths killing over 40,000 under-five children every year [11]. These deaths are easily preventable and treatable through simple and cost effective interventions. Immunization, good nutrition, exclusive breast feeding, appropriate complementary feeding and hand washing are among the preventive while administration of amoxicillin dispersible tablets and other antibiotics are among the curative methods which can save lives.

With the objective of increasing access to these lifesaving interventions, Ethiopia has made a policy breakthrough of introducing community based treatment of pneumonia through health extension workers in 2010 [11]. Since then over 38,000 health extension workers from nearly 15,000 health posts are equipped with the skills and supplies to treat pneumonia at community level using the integrated community case management approach [11]. Early diagnosis and treatment of pneumonia, and access to health care, will save lives, thus strategies must target low income communities.

1.2 STATEMENT OF THE PROBLEM

Retrieving relevant solution for the user problem is milestone in reasoning systems. The output of the case based system will be in the way that not in dilemma for the user. Solution that will be displayed should be accurate. Retrieving the most similar cases from the case base needs good representation format of cases including how much relevant the cases are in representing the domain area and strong retrieval algorithm.

Slow response time in retrieving the best similar case from the case base is also one of the problems of case based reasoning system. As research indicates this is occurred because of the number of cases is increased from time to time since the CBR learns from the user cases. So to retrieve the best case it will take more and more time whenever the numbers of cases are increased.

The third most important problem is flourished from the broad domain area which is medical disease diagnosis. After made a survey on the broad domain area, we have got a problem on the

pneumonia disease which later became our specific domain area. The problem is that pneumonia is the most killing disease in Africa especially in Ethiopia [7] and there is lack of experts and generally professionals on health area. Shortage of doctors in developing countries is became one of the cause for death for many peoples. The ratio of doctors to patient in developing counties is very high [9]. The problem indicate the need for more medical doctors and more expert system that can minimize the gap between doctors and patients.

The research questions derived from the statement of problem for this research in titled "An Integration of Case Based Reasoning and Data mining Techniques to Enhance the Effectiveness: In Case of Pneumonia Disease Diagnosis" are the following:

1. How can we improve retrieval performance through more effective approaches?
2. How to identify cases and rules using data mining technique and provide to CBR system instead of human experts?
3. How we combine general models and specific case experiences to achieve the intended support?
4. What is the significance (i.e. usefulness, impact) of such a system in clinical decision making?

1.3 SIGNIFICANCE OF THE STUDY

The main purpose of this thesis are discussed in the following three points.

- The aim is to create combined (CBR paradigm with DM techniques) formalisms that benefit from each of their components. The disadvantages or limitations of specific intelligent methods can be surpassed or alleviated by their combination with other methods.
- In medical diagnosis task strong reasoning is need to prescribe the treatment for the patient because it is human life at risk. We are using two kinds of knowledge in our prototype system which is specific case experienced knowledge and general knowledge from the dataset generated in the form of rule. The hybrid reason to be strong enough to diagnosis and recommends the treatment.
- To introduce and provide good result by Integrating CBR and DM technique.

1.4 MOTIVATION

The motivation of this studying on the integration of CBR and DM techniques for pneumonia disease diagnosis are explain in to two perspectives, which are from medical or health point of view and from computer reasoning techniques point of view.

Disease diagnosis Case are collected from hospital medical records and standard guidelines, if the case data are available, it is not expensive and time consuming compared to rule based reasoning which use different data collection and Knowledge acquisition techniques. Probably when there are missing or incomplete data in CBR system [10], it works well even if the similarity is low. This is because of CBR inference engine does not need the problem part to be 100% similar to retrieve the solution part. As we used CBR methodology in real life at many places it is not difficult to convince the developers, user and managers of the validity of the paradigm. In our daily life activities, we solve problems in a CBR paradigm fashion by recalling how we or someone else solve similar problems in the past. The hybrid paradigm has many importance as compared to using the standalone single techniques. The hybrid of CBR and DM techniques has many benefits that inspire researchers on hybrid artificial intelligence techniques.

There are two main knowledge categories which are general and specific knowledge [14]. The CBR paradigm represents specific knowledge type and the DM techniques used to formulate rules that represent general knowledge. The hybrid reasoning techniques and knowledge representation formalism benefit from the two forms of knowledge and reasoning techniques.

When we look from the medical point of view, compared to other health problems pneumonia disease is the most killing disease [24], that needs more focus and intelligent application. In Ethiopia doctor to patient ratio is high compare to other countries which is 1 doctor to 35,000 patient [9], so we need to have more intelligent application systems. On the other hand pneumonia is more a symptom based disease to diagnosis. It means that less laboratory test, more symptom and sign identification are used to diagnosis pneumonia. It is easy for users of intelligent application to be diagnosis as pneumonia patient or not.

1.5 OBJECTIVES OF THE STUDY

The general and specific objectives of the study are described in the following two sections as indicated below.

1.5.1 GENERAL OBJECTIVE

The general objective of the research is to enhance the effectiveness of pneumonia disease diagnosis by developing prototype system using integration of case based reasoning and data mining techniques.

1.5.2 SPECIFIC OBJECTIVES

The specific objectives of the research are mentioned below.

- To understand the concept of case based reasoing system and data mining techniques and how it is designed by reviewing different literatures.
- To acquire pneumonia disease knowledge from different sources like experts and documents and model the acquired knowledge using appropriate modeling technique.
- To design and develop a prototype system using case based reasoning and data mining techniques.
- To evaluate and test the efficiency of the developed prototype.
- To reach conclusion and recommendation from the research conducted.

1.6 CONCEPTUAL FRAMEWORK CBR AND DATA MINING

The conceptual frame work for integrating data mining techniques and case based reasoning are discussed below.

1.6.1 CASE BASED REASONING

CBR as a field deals with both theoretical and practical problems to convey reliable and plausible reasoning. It has been affected by different fields, for instance, cognitive science, knowledge-based systems, machine learning, databases, information retrieval, fuzzy logic and neural networks.

There are also a lot of commonalities with other fields such as uncertainty, pattern recognition, and statistics [12].

Central tasks that all CBR methods have to deal with are "*to identify the current Problem situation, find a past case similar to the new one, use that case to suggest a solution to the current problem, evaluate the proposed solution, and update the system by learning from this experience"* [2]. All cases stores in the case base probably contain two parts which is a problem description and its related solution. Based on Aamdot and plaza (1994) all CBR tasks have 4 phases and denoted by 4RE which is retrieval, reuse, revision and retain. Later it is extended by adding represent step by Finnie and Sun in 2003 which is said to be 5RE steps.

The basic assumption behind CBR is that problems have a way of reoccurring in a similar way. It means that similar problems will have similar solution. The solution we applied last time, may therefore also apply for this new but quite similar problem. Take for instance a doctor examining a patient with a set of given symptoms. While listing the symptoms the doctor is reminded of a previous case, a patient with the same symptoms who came in a couple of weeks ago. The patient was treated with antibiotics and told to hold the bed, and he recovered in no time. Might the same treatment work in this instance?

1.6.2 DATA MINING TECHNIQUES

The study of foundations of data mining should be viewed as a scientific inquiry into the nature of data mining and the scope of data mining methods. The study of the nature of data mining concerns the philosophical, theoretical and mathematical foundations of data mining; while the study of data mining methods concerns its technological foundations by focusing on the algorithms and tools [13].

Data mining, as a relatively new branch of computer science, has received much attention. It is motivated by our desire of obtaining knowledge from huge datasets. The knowledge obtained using different data mining techniques can be used for different knowledge oriented application. Many data mining methods, based on the extensions, combinations, and adaptation of machine learning algorithms, statistical methods, relational database concepts, and other data analysis techniques, have been proposed and studied for knowledge extraction and abstraction [13].

1.7 SCOPE AND LIMITATION OF THE STUDY

The Scope and Limitations of the study described below.

1.7.1 SCOPE

- The domain area used is pneumonia diseases diagnosis using CBR and DM techniques: So that our research is not concerned on any other disease including lower respiratory tracts infections.
- The thesis includes the implementation of a prototype system with basic CBR and DM techniques: The CBR system will be limited to the case description and case solution and does not implement solution justification and Result.
- The DM will be a fully functional process which uses CRISP-DM data mining process cycle but instead of using DM algorithms from the existing tool randomly, we surveyed the best performed algorithms and implemented them in the system.
- The report provides a thorough discussion of the implementation and the pros and cons of the implementation: There is also an evaluation of the results found by the explanation mechanism implementation, but no in-depth discussion of the system implementation itself, i.e. the combination of CBR and DM.

1.7.2 LIMITATIONS OF THE STUDY

- Most clinical guidelines are not developed in a format that allows for straightforward incorporation into computerized clinical decision support systems.
- Case data is more complex, and this makes feature mining more complex.
- There is lack of open source pneumonia data sets, tools and frameworks.

1.8 BENEFICIARIES OF THE STUDY

The study will have the following beneficiaries

- User to get approximate right answer towards their pneumonia related health problems.
- Using case based reasoning and data mining techniques to search and provide information about pneumonia disease including the Pneumonia type and treatments.
- To train primary healthcare workers, general practitioner.

1.9 ORGANIZATION OF THE THESIS

Chapter One: we look at some background information, both related to pneumonia disease and to the different approaches we aim to use for the research.

Chapter Two: we assess some of the related research within disease diagnosis using different Reasoning system. The chapter has, two main different parts literature review which contains the general concepts, definitions and categories of DM and CBR, and related works which has three sub parts of research paper reviewing: CBR, DM Techniques and Integration of CBR and DMT.

Chapter Three: we describes the approaches and methods used during the project including Data Collection and prototype design process.

Chapter Four: we propose the initial architecture for integration of CBR and DM Techniques. Preprocessing of pneumonia dataset using a weka tool is held here. How we apply the CRISP-DM process is explained in this chapter.

Chapter Five: Implementation of the integrated architecture.

Chapter Six: describes the evaluation, result and discussion part of the research.

Chapter Seven: describes summery of findings, conclusions and recommendations.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

This chapters provides literature Review and related work on CBR and DM techniques.

2.1 LITERATURE REVIEW

This section provides an explanation about the basic concepts, definitions and approaches that are related to our studies.

2.1.1 KNOWLEDGE BASED SYSTEMS

The concept of knowledge based systems is derived from the field of artificial intelligence (AI). AI intends understanding of human intelligence and building of computer programs that are capable of simulating or acting one or more of intelligent behaviors. Intelligent behaviors include cognitive skills like thinking, problem solving, learning, understanding, emotions, consciousness, intuition and creativity, language capacity, etc. These days some of the behaviors such as problem solving, learning and understanding are handled by computer programs [14].

Computer programs that solve problems in a human expert-like fashion by using knowledge about the application domain and problem solving techniques are known as Knowledge based system. Knowledge based systems handle problems in the same way like human experts. They represent the knowledge about the application domain and they use one or more techniques that guides on how to use the knowledge to solve problems. Every knowledge based system has two main components which are known as knowledge base and inference engine [14].

The knowledge base contains all relevant knowledge about the domain area. The knowledge can be extracted from different sources such as experts, documents, books and/or other sources. There are different formats to represent knowledge in the knowledge base. The two common ways to represent knowledge are cases and rules.

The first format to represent the knowledge is a case. "*A case is a contextualized piece of knowledge representing an experience that teaches a lesson fundamental to achieving the goals of*

the reasoner [15]." The case represents specific experienced knowledge. The experience a case represents can be structured in various ways.

Very often it is only subdivided into a problem and a solution description, but additional knowledge might be necessary depending on the kind of intended reuse. For example, (Kolodner,1993) proposes a comprehensive case structure consisting of the following five parts: (i) a situation and its goal, (ii) the solution and, sometimes, means of deriving it, (iii) the result of carrying it out, (iv) explanations of results, and (v) lessons that can be learned from the experience [15].

Rules are the second common format to represent knowledge in the knowledge base. Rules represent what to do or not to do while certain situations are met. Similarly, the application domain knowledge is represented with set of rules that represent the facts that would be true when some conditions are given true. The basic form of rule is If <conditions> then <conclusion>. The conditions of a rule are connected between each other with logical connectives such as AND, OR, NOT etc. thus forming a logical function. When sufficient conditions of a rule are satisfied, the conclusion is derived and the rule is said to fire (or trigger). Rules represent general knowledge regarding a domain [10].

The second component of a knowledge base system is inference engine. After the system gets the required knowledge, it needs to be instructed how to use the knowledge in solving problems. Inference engine represents the reasoning technique that manipulates, uses and controls the knowledge to solve the problems. Case based reasoning and rule based reasoning are two examples of reasoning techniques [14].

2.1.1.1 CASE BASED REASONING TECHNIQUE

Case based reasoning, as its name indicates, use cases to reason about a given problem. In its problem solving process, it reuses old similar cases to understand the problem, suggest a solution, and/or to keep it from failure. A case based reasoning technique follows four processes; retrieve, reuse, revise and retain, to accomplish its reasoning task according to A. Aamodt and E. Plaza (1994) [2], But Finnie and Sun (2003) add one important phase which is represent. knowledge should be represented before retrieved. There are five phases of CBR process. Figure 1 shows sequence of the processes and each process is described below.

I Represent

The case in the CBR system mainly represented in to two parts which are problem description and solution description [14].

- ***Situation/Problem description****:* describes specific circumstances, states of a situation, and state of the environment when particular case is recorded.
- ***Solution****:* provides how the problem described in the problem description was solved or treated in a particular instance.

II Retrieve

An initial description of a problem is get from the user of the system which is a *new case*. Based on a new case the system will RETRIEVE a case from the collection of previous cases. A new user problem is solved by ***retrieving*** one or more previously experienced cases.

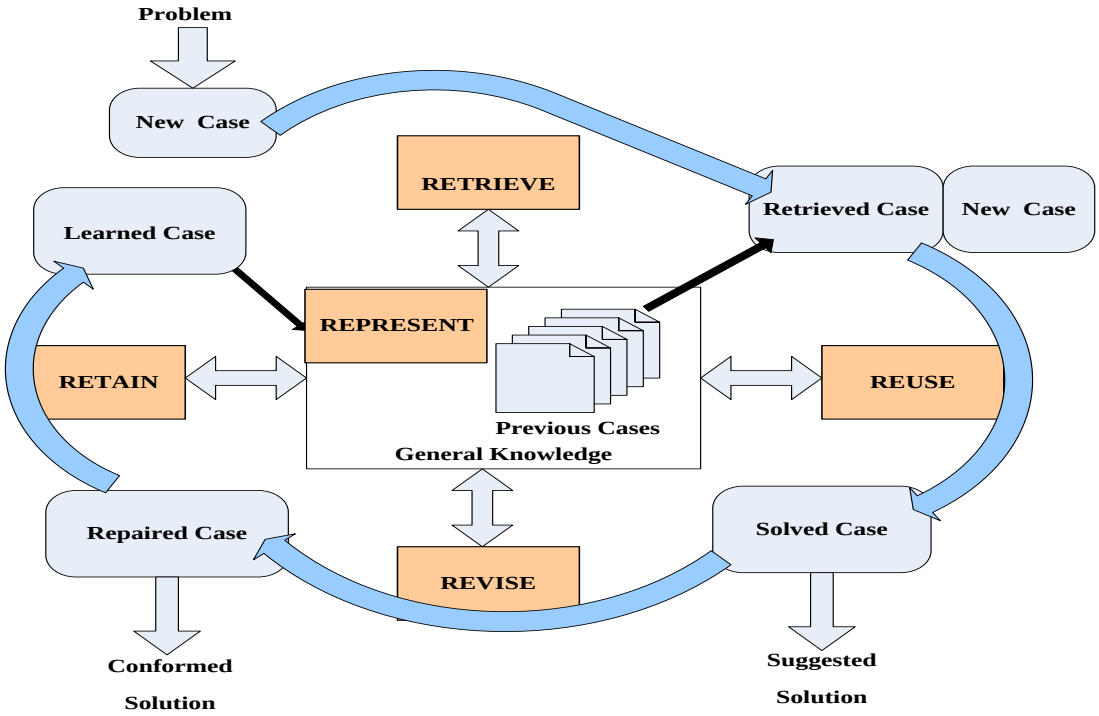

Figure 1: CBR Process Cycle Adapted from [2]

III Reuse

The retrieved case is combined with the new case - through REUSE - into a solved case, i.e. a proposed solution to the initial problem. The reuse of the solution can be done in two ways.

One is just copying the solution of the retrieved case as the solution for the target case (null adaptation). This is applicable to classification applications. However, in most applications, a retrieved solution cannot be used directly as the solution of the target case and some adaptation is necessary [16].

IV Revise

Revision can be viewed as two tasks: diagnosis of failure and solution repair [17]. For diagnosis, one of the following ways can be used: 1) Execution of the solution and evaluating the outcome. 2) Using a simulation -model of the real world and evaluate the solution using the model. This solution is safer and more cost effective. 3) Experts also can help in diagnosing the failure in solutions. The expert evaluates the solution using his/her experiences. 4) Use the case base itself to identify the failure. In this case, to assist in problem diagnosis, in addition to the specification of the problem and the solution for each case, knowledge about the conditions under which the failure may occur must also be stored in the case base [16].

V Retain

In a CBR system learning is done in the retention step. In this step, the new case will be added to the case base according to some policies in the system. Retention includes adding knowledge and new cases to the case base, all which needs to be indexed, as well as deleting cases from the case base in order to restrict its growth. Having new information about the cases in the case base and the knowledge system obtained in the previous steps in the cycle, indexing of the case base and other knowledge would be changed in this step. Different retention strategies fall into one of two groups: maintenance of the content of the case base and maintenance of the organization of the case base. Research on maintaining the content involves work on the reduction of the case base and the deletion and addition policies. Maintenance of the organization is related to indexing the case base in order to make the case retrieval faster and more efficient [16].

Advantage and Disadvantage of CBR

We describe some of its advantages and disadvantages from various points of view [10].

Advantage of CBR

- ***Ability to express specialized knowledge:*** This feature of cases among other advantages circumvents interpretation problems suffered by rules due to their generality.
- ***Modularity***: Each case is a discrete, independent knowledge unit that can be inserted into or removed from the case base, without any problem.
- ***Easy knowledge acquisition***: Knowledge acquisition in case-based representations is not usually a problem, due to the fact that cases are available in most application domains. However, there are domains where they are not.
- ***Handling unexpected or missing inputs***: can handle unexpected cases not recorded in the system or missing input values by assessing their similarity to stored cases and reusing relevant cases.
- ***Reflecting human reasoning***: used CBR methodology in real life at many places it is not difficult to convince the developers, user and managers of the validity of the paradigm.
- ***Inference efficiency***: Adapting preexisting cases to handle new problems is usually more efficient than having to solve a problem from scratch as in rule-based systems. This is not always true look for disadvantage of CBR.
- ***Extending to a broad range of domains***: This is due to the appearing of unlimited number of ways representing, indexing, adapting and retrieving cases.

Disadvantage of CBR

- ***Inability to express general knowledge***
- ***Knowledge acquisition problems***: Although knowledge acquisition is not a problem when a sufficient number of cases are available in a domain, various knowledge acquisition problems may arise when dealing with domains where cases either are unavailable or are available in a limited amount.
- ***Inference efficiency problems***: Degradation of the time efficiency of case retrieval is associated with the utility problem, a problem occurring in learners when knowledge learned in an attempt to improve a system's performance degrades performance instead.

- ***Provision of explanations***: Some kind of explanation can be provided for the conclusions reached, but not in a straightforward manner as in rule-based systems.

2.1.1.2 RULE BASED REASONING

Rule based systems are knowledge based systems that represent the domain knowledge with set of rules and suggest a solution or conclusion of a problem by using rule based reasoning method. A rule based system has three components, which are known as working memory, knowledge base (Rule Base) and inference engine [14], [18].

Inference Engine receives a problem from the working memory and provides the reasoning result to the working memory. The inference engine contains three sub parts which are pattern matcher, An Agenda, An execution engine. The details of how RBR works are described in in figure 2 below.

- ***Pattern Matcher:*** All the rules are compared to working memory using the pattern matcher to decide which ones should be activated during this cycle. This unordered list of activated rules, together with any other rules activated in previous cycles, is called the conflict set.
- ***An Agenda (conflict set):*** the list of rules whose right-hand sides will be executed, or fired. The process of ordering the agenda is called conflict resolution. The conflict resolution strategy for a given rule engine will depend on many factors, only some of which will be under the programmer's control.
- ***An Execution Engine***: the first rule on the agenda is fired (possibly changing the working memory) and the entire process is repeated.

Working Memory contains the description of the problem and updates its content based on the reasoning results received from the inference engine.

***Rule Base* (*knowledge base*):** contains all the rules the system knows. They may simply be stored as strings of text, but most often a rule compiler processes them into some form that the inference engine can work with more efficiently. Some rule engines allow (or require) you to store the rule base in an external relational database, and others have an integrated rule base. Storing rules in a relational database allows you to select rules to be included in a system based on criteria like date, time, and user access rights [18].

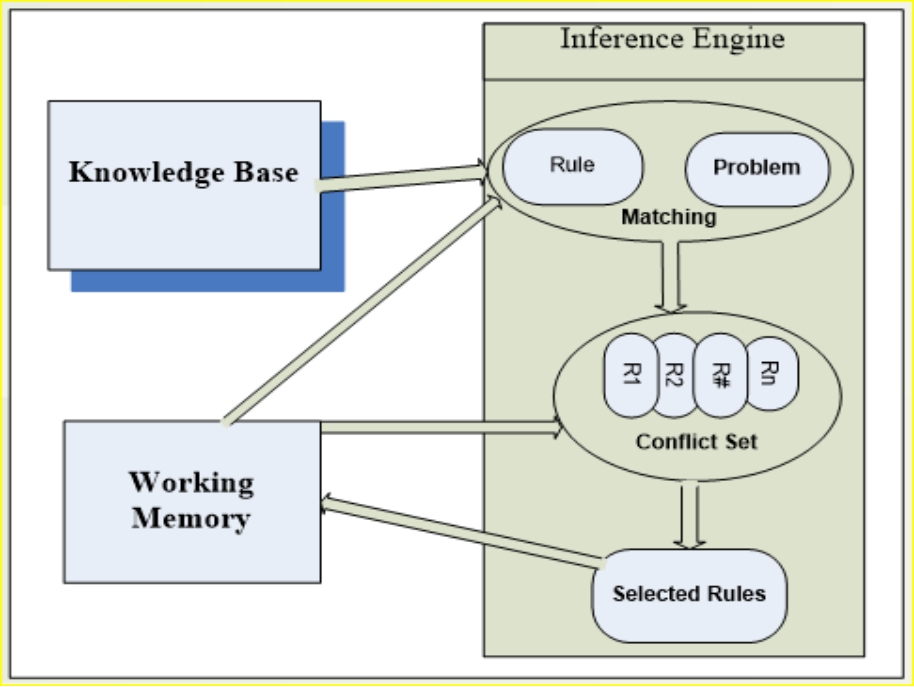

Figure 2: Rule Based Reasoning [14]

Rule based reasoning technique represents how a system solves a problem by using knowledge of the application domain that is represented in form of rules. There are two ways of rule based reasoning methods: forward chaining and backward chaining.

- ***Forward-chaining rule*:** is something like an if ... then statement in a procedural language, but it is not used in a procedural way. Whereas if ... then statements are executed at a specific time and in a specific order, according to how the programmer writes them.
- ***Backward-chaining system*:** rules are still if … then statements, but the engine actively tries to make rules fire. If the if clause of one rule is only partially matched and the engine can determine that firing some other rule would cause it to be fully matched, the engine tries to fire that second rule. This behavior is often called goal seeking.

Advantage and Disadvantage of Rule based Reasoning

The following are some advantages and disadvantages of RBR from various points of view.

Advantage of RBR

A RBR approach has tremendous advantages in the development of knowledge based system. The following are the main advantages of RBR [10].

- ***Compact representation of general knowledge***: Rules can easily represent general knowledge about a problem domain in autonomous, relatively small chunks.
- ***Naturalness of representation***: Rules are a very natural knowledge representation method, with a high level of comprehensibility, since they look like natural language expressions.
- ***Modularity***: Each rule is a discrete knowledge unit that can be inserted into or removed from the knowledge base, without having to take care of any other technical detail.
- ***Provision of explanations***: The ability to provide explanations for the derived conclusions in a straightforward manner is a vital feature, given that explanations in certain application domains (e.g. medicine) are considered necessary.

Disadvantage of RBR

The drawback or problems of RBR that makes us not confidentially or comfortably use it are described below [10].

- ***Knowledge acquisition bottleneck***: The standard way of acquiring rules through interviews with experts is cumbersome and time consuming.
- ***Brittleness of rules***: It is not possible to draw conclusions from rules when there are missing values in the input data.
- ***Inference efficiency problems***: Inference handles each problem from scratch, whether it has dealt with the same problem successfully in the past or not. This creates inefficiency, especially in cases when the problem-solving process is time consuming.
- ***Difficulty in maintenance of large rule bases***: Whenever the number of rules increases, the rule base may have problems such as redundant rules, conflicting rules, rules with redundant or missing conditions, missing rules etc.
- ***Problem-solving experience is not exploited***: A rule-based system is not self-updatable, in the sense that there is no inherent mechanism to incorporate experience acquired from dealing with past problems
- ***Interpretation problems***: The general nature of rules may create problems in the interpretation of their scope during reasoning. To effectively deal with a specific situation, rules may sometimes need to be specialized.

2.1.1.3 HYBRID REASONING TECHNIQUES

The combination of two or more different problem-solving and knowledge representation methods is a very active research area in artificial intelligence. The aim is to create combined formalisms that benefit from each of their components. It is generally believed that complex problems are easier to solve with hybrid or integrated approaches. There are mainly three kinds of hybrid(Coupling) Approaches explained in [10].

- ***Sequential Processing:*** The sequential processing category refers to coupling approaches in which the flow of information between the integrated modules is sequential or semi-sequential. It includes approaches in which information necessarily passes sequentially through some or all of the integrated components in order to produce a final result. We distinguish two subcategories of the sequential processing category, based on the importance of the information (besides the input case) passed between the coupled modules.

  The Loosely Coupled Sequence: subcategory refers to approaches in which the reasoning process of each component is almost independent of the reasoning process of the previous component in the sequence. The output of one component may be used as input knowledge to the next component but does not play an important role in the internal reasoning process of the next component; it is mainly used as triggering information.

  The Tightly Coupled Sequence: subcategory refers to approaches in which the information passing to a component from a previous one in the sequence plays an important role in its reasoning process. Approaches belonging to this subcategory are more closely coupled than approaches in the loosely coupled sequence subcategory.
- ***Coprocessing:*** refers to approaches in which the components closely interact (as partners) in producing the final result. In such approaches, information flow between the components is bidirectional. The integrated components may also work in parallel for the solution of a problem. Systems belonging to this approach are discerned into two types: cooperation oriented, which gives emphasis on cooperation, and reconciliation oriented, which gives emphasis on reconciliation. In the former type, the integrated components cooperate with each other (usually by interleaving their reasoning steps) for the production of the final result.

- ***Embedded processing:*** In embedded processing approaches, a component based on one representation method is the primary problem solver, embedding one or more components based on the other representation method to handle its internal reasoning tasks.

Summarized Review of Related Studies

The effectiveness of various hybrid or integrated approaches has been demonstrated in a number of application areas. Some of the popular types of combinations are discussed below.

I Integration of CBR and RBR

The integration of CBR with different RBR methods are discussed in the table 1 below.

Table 1: Integration of CBR with different RBR methods

Research Title	***Application Domain***	***Method***	***Result and Evaluation***	***Reference***
A Hybrid Recommender System Using Rule-Based and Case-Based Reasoning	Movies Rating	It produces recommendations by applying two model-based collaborative filtering (CF) techniques. K-means clustering algorithm to classify users into different clusters. These clusters are used as a basis to select a neighborhood of an active user. Consequently, for the process of neighborhood selection, proposed system utilizes two knowledge-based techniques: RBR and CBR separately.	Use two metrics to assesse the accuracy of CF. prediction accuracy metrics and classification accuracy metrics.	[19]

Integration of RBR Expert Systems and CBR in an Acute Bacterial Meningitis Clinical Decision Support System	Acute Bacterial Meningitis	Focused on the implementation of the adaptation stage, from the integration of CBR and Rule Based Expert Systems. The framework is initially applied to the pre-diagnosis stage which uses a set of basic diagnostic rules so as to identify situations in which a certain set of symptoms indicate the presence of a disease without a shadow of a doubt. The retrieval is implemented by the nearest neighbor method. The local similarity between symptoms is determined by equality.	The experiment use 30 case sample. This diagnosis was later compared to the one provided by the expert to determine whether the system hit or failed. This experiment shown an accuracy of 97%.	[20]

Integration of case-based and rule-based reasoning through fuzzy inference in decision support systems	They use IRIS data set for experimental purpose.	In Stage I, Cases can be gathered in the case base for further using in decision-making. In Stage II DM methods used to automatically, extract new knowledge in the form of rules from cases sample. In Stage III, the transition to a rule-based model of knowledge representation means that we obtain the explicit form of knowledge, able to explain the causal relationships. This explicit knowledge can be cleaned, refined and interpreted by experts. In Stage IV, the accumulation of new cases takes place using the received version of a decision support system.	To resolve conflicts between the rules the 1st strategy was used for 3MF, 5MF & 7MF. Comparison B/N 2 strategies of conflict resolution shown for 3MF & 5MF. The best accuracy on average is achieved for 3MF, which is 0.860 for conflict resolution 1 and 0.939 for conflict resolution 2.	[21]

II Integration of CBR With Soft Computing

According to [12], Soft computing is a field of computer science that studies the possibility of finding new models to deal with cognitive functions' problems. Such problems can cover but are not limited to perception, systematic thinking, reasoning, object recognition, data mining, episodic memory, control, and knowledge management. The techniques that are normally utilized to establish such models are FL, ANN, PR and GA.

Table 2: CBR With Soft Computing Using Different Method

Research Title	***Application Domain***	***Method***	***Result and Evaluation***	***Reference***
CBR with Bayesian Model Averaging: An Improved Method for Survival Analysis on Microarray Data	Microarray Data: gene expressions	The first component of the system is Bayesian Model Averaging (BMA) feature selection system serve as indexing mechanisms to search through the case memory. The posterior probabilities provide weights for calculating the similarity measure. The second component of the system is CBR step		[22]
A fuzzy-ontology-oriented CBR framework for semantic diabetes diagnosis	diabetes diagnosis	It proposes a fuzzy case-based OWL2 ontology, and a fuzzy semantic retrieval algorithm	Accuracy of 97.67%	[23]

Combining CBR with GA optimization for preliminary cost estimation in construction industry	Cost estimation in construction industry	The CBR framework used for generating a CBR application for modelling the new EVAS-CBR system, database management framework used for manipulating HRB projects that constitute a case library, and GA framework used for optimizing Cost Factor weights. System frameworks play an important role in the selection of an appropriate case from the case library.	590 real cases used 570 cases to construct the case library and 20 cases for system validation. The error rate of the GA-based CBR system was 5.60%. This rate is acceptable considering that the accuracy range of the error rate of American Association of Cost Engineers (AACE) is in the range of –10% to + 15%.	[24]

III Integration of RBR With Soft Computing Methodology

The integration of CBR with different soft computing methods are discussed in the table 2 below.

Table 3: RBR With Soft Computing Using Different Method

Research Title	***Domain Area***	***Method***	***Result and Evaluation***	***Reference***
Connectionist Expert System for Medical Diagnosis using ANN- A case study of skin disease Scabies.	Skin disease Scabies	The system is implemented using Mat lab and a feed forward multilayer network is used. Back propagation algorithm is used for training the network.	It achieves 95% success.	[25]

An Hybrid Architecture Integrating Forward Rules with Fuzzy Ontological Reasoning	Decision-support system for the tourism domain	The Drools framework (language and engine) to support uncertainty reasoning upon rules, they have integrated it with custom operators that exploit Pellet to perform ontological reasoning, and exploit FuzzyDL to support fuzzy ontological reasoning.		[26]
Multi-Modal Reasoning Medical Diagnosis System Integrated With Probabilistic Reasoning	Chinese medicine diagnosis	The system use CBR and RBR with Bayesian Networks (BN). BN for Probabilistic Reasoning	The average accuracy of disease diagnosis (89%) is much better than the average accuracy of syndrome diagnosis (47%). Usage percentage of patient records for diagnosis is 62%.	[27]

2.1.2 DATA MINING TECHNIQUES

The study of foundations of data mining should be viewed as a scientific inquiry into the nature of data mining and the scope of data mining methods. The study of the nature of data mining concerns the philosophical, theoretical and mathematical foundations of data mining; while the study of data mining methods concerns its technological foundations by focusing on the algorithms and tools. The existing studies of data mining can be classified broadly under three views [13].

2.1.2.1 FUNCTION-ORIENTED VIEW

Focuses on the goal or functionality of a data mining system, namely, the discovery of knowledge from data. In a well-accepted definition, data mining is defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data". Depending on the data and their properties, one may consider different data mining systems, functionalities and for different purposes, such as text mining, Web mining. Under the function-oriented view, the objectives of data mining can be divided into prediction and description. *Prediction* involves the use of some variables to predict the values of some other variables, and *description* focuses on patterns that describe the data. Some functionalities are clearly well defined and researched, among some are listed and described [28], [29] below.

- ***Classification/prediction:*** classification is a supervised data mining method applied to datasets containing an expert labeling in the form of a categorical attribute, called a class; when the attribute is numeric, the method is called prediction. Examples of classifiers include neural networks, support vector machines (SVMs), naïve Bayes, and decision trees.
- ***Association Mining:*** association mining mines for frequent itemsets in a dataset, which can be represented as rules such as in market basket analysis. It is an unsupervised method. The most famous algorithm in this category is apriori algorithm.
- ***Clustering:*** clustering finds groups of similar objects in a dataset, which are also dissimilar from the objects in other clusters. In addition to the similarity-based methods like K Means, some methods use density-based algorithms or hierarchical algorithms. Considerations for evaluating the mining results vary in these different methods, however a set of quality measurements are traditionally associated with each, for example accuracy or error rate for classification, and lift or confidence for association mining.

2.1.2.2 THEORY-ORIENTED VIEW

Any theory discovered and used by scientists can be used by data mining systems. Thus, many fields contribute to the theoretical study of data mining. They include statistics, machine learning, databases, pattern recognition, visualization, and many other. There is also a need for the combination of existing theories. For example, some efforts have been made to bring the rough

sets theory, fuzzy logic, utility and measurement theory, concept lattice and knowledge structure, and other mathematical and logical models into the data mining models.

2.1.2.3 PROCESS-ORIENTED VIEW

From the procedure (process-oriented) view, data mining deals with a "non-trivial" process consisting of many steps, such as data selection, data preprocessing, data transformation, pattern discovery, pattern evaluation, and result explanations [28]. Furthermore, it should be a dynamically organized process. Under the process-oriented view, data mining studies have been focused on algorithms and methodologies for different processes, speeding up existing algorithms, and evaluation of discovered knowledge.

There are mainly three kinds of data mining process methodologies are available [28]. These are The SEMMA Analysis Cycle for SAS Enterprise Miner, the 5A's Process for SPSS Clementine, and the CRISP-DM.

- ***SEMMA Analysis Cycle for SAS Enterprise Miner:*** which have five steps
  - ***Sample***: To create one or more data tables by sampling data from data warehouse.
  - ***Explore***: To explore the data visually or numerically for trends or groupings.
  - ***Modify***: Refers to creating, selecting, and transforming one or more variables to focus the model selection process.
  - ***Model***: Creating a data model involves using the data mining software to search automatically for a combination of data that predicts the desired outcome reliably.
  - ***Assess***: A model is to set aside a portion of the data during the sampling stage. If the model is valid, it should work for both the reserved sample and for the sample that was used to develop the model.
- ***SPSS Clementine 5A's Process:*** contains five data mining analysis cycle which are Assess, Act, Access, Automate and Analyze. During the second quarter of 2001, SPSS removed all references to the 5 A's process methodology from its web site. SPSS now actively supports the CRISP-DM process model.
- ***CRISP-DM:*** is an industry- and tool-neutral data mining process model that which would apply on small as well as large data mining projects faster, cheaper, more reliable and more manageable. It is the most popular knowledge discovery process model. In CRISP-DM the

numbers of steps followed are six and some of the steps are iterative. Each step has good documentation and divided into sub steps which help easily to identify all necessary details in the knowledge discovery process [28], [29].

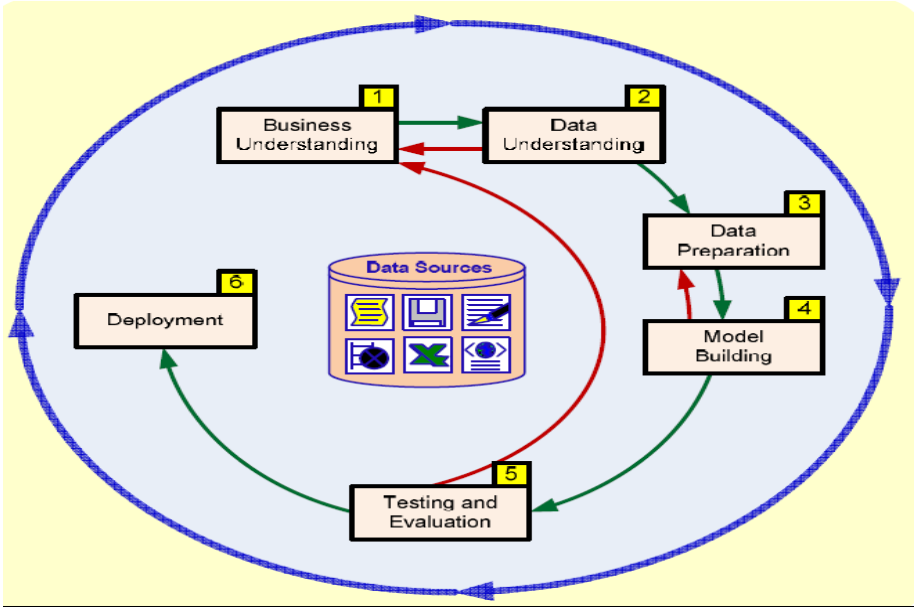

Figure 3: CRISP-DM Process Model [29]

Business Understanding: Understanding the project objective and requirement from the business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

Data Understanding: To identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation: The process of constructing final dataset using preprocessing tools such as weka. data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling: various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

Evaluation: model (or models) built appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

Deployment: The knowledge gained will need to be organized and presented in a way that the client can use. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. The sub processes of the six steps of CRISP-DM are explain in the Integration Architecture chapter. The details of how we use CRISP-DM process model in this research will be seen in chapter 4.

2.1.3 KNOWLEDGE ENGINEERING PROCESS

According to [35] the process of developing models, methods, and basic technologies for acquiring and representing knowledge and for building intelligent knowledge-based systems, is called knowledge engineering. Knowledge acquisition and representation are discussed in the section below.

2.1.3.1 KNOWLEDGE ACQUISITION

Knowledge acquisition is the process of gathering information from the field and using it in some way. Typically it involves retrieving information from one to several experts and/or documentation, representing it in some way and translating it into a way that machines can understand. In this research mainly we use document as a source of knowledge. The document is the Patient Medical record cards which contains patient diagnosis information [31].

We use KDD (Knowledge Discovery in Databases) process to extract relevant data from a given database and apply common machine learning methods with the help of the data mining software. The process was in-depth and time-consuming which is what one might expect from learning something from a new field.

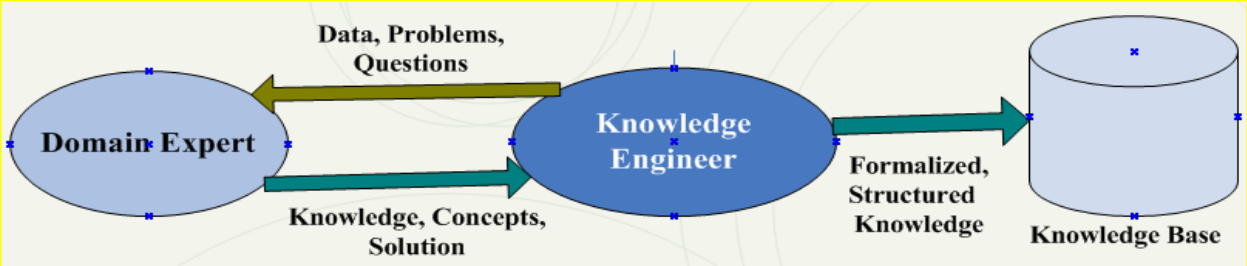

Figure 4: Simple Knowledge Acquisition Process [31]

A common way to retrieve knowledge is to hold interviews with the experts, discussing the different aspects of the field or presenting cases to see how and what the expert emphasizes in the problem solving. In relation to presenting cases, it is also possible to observe the experts working in their natural habitat, to see if there are things they do that may be important, but they are not aware of them self. Figure 4 above illustrates a simple knowledge acquisition process where a knowledge engineer queries the domain expert and formalizes the information into structured knowledge.

2.1.3.2 KNOWLEDGE REPRESENTATION

Knowledge can be represented on different format for different reasoning techniques. Knowledge representation for CBR and RBR are in different formalism. Knowledge representation for RBR and DM are different to some extent.

I Knowledge Representation for CBR

As we know there are different knowledge representation format for CBR system. The two commonly used knowledge representation formalisms are discussed below in [15].

- ***Feature-vector approaches***: represent a case as a vector of attribute-value 0pairs, similar to the propositional representations used in Machine Learning (ML), that support k-nearest neighbor matching and instance-based learning.
- ***Structured approach***: to case representation originates from the episodic memory notion of cases. The representational structure itself is usually developed around a frame-based formalism. Since frames can be seen as a subset of first-order logic, cases represented as frames or some frame-like structure are examples of what ML calls relational representations. Such representations arise from gathering together clusters of relations that occur together in elementary objects. Cases, from this perspective, are clusters of relations

between the kinds of elementary objects that comprise them. Frame representations also strongly resemble object-oriented case representations [15].

II Knowledge Representation for DM Technique

There are different formats of knowledge representation for input and output of DM process. Some of those format are listed and described in [32] below.

- ***Table***: is the simplest, most rudimentary way of representing the output from machine learning is to make it just the same as the input table.
- ***Tree***: is a divide-and-conquer approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree. Nodes in a decision tree involve testing a particular attribute. Usually, the test compares an attribute value with a constant. Leaf nodes give a classification that applies to all instances that reach the leaf, or a set of classifications, or a probability distribution over all possible classifications. To classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf.
- ***Rules***: are a popular alternative to decision trees. The antecedent, or precondition, of a rule is a series of tests just like the tests at nodes in decision trees, while the consequent, or conclusion, gives the class or classes that apply to instances covered by that rule, or perhaps gives a probability distribution over the classes. Generally, the preconditions are logically ANDed together, and all the tests must succeed if the rule is to fire.

  Classification Rules: are easy to read a set of classification rules directly off a decision tree. One rule is generated for each leaf. The antecedent of the rule includes a condition for every node on the path from the root to that leaf, and the consequent of the rule is the class assigned by the leaf. This procedure produces rules that are unambiguous in that the order in which they are executed is irrelevant.

  Association rules: are no different from classification rules except that they can predict any attribute, not just the class, and this gives them the freedom to predict combinations of attributes too. Also, association rules are not intended to be used together as a set, as

classification rules are. Different association rules express different regularities that underlie the dataset, and they generally predict different things.

2.1.4 KNOWLEDGE VALIDATION

As [33] mentioned, there are four types of evaluations to be conducted on KBS. These are verification, validation, usability and usefulness.

Verification: is the rightness of the developed KBS to be evaluated. It can be conducted entirely on the formal model or on the computable model whose syntax is clearly stated for their rightness to be evaluated. It assures whether the knowledge on the formal model or on the computable model does not comprise syntactical faults. A verified KBS denotes the acquired knowledge from domain experts and secondary sources rightly.

Validation: is checking the knowledge base of the KBS for semantic faults that may occur during the KBS development. A validated KBS comprises the correct knowledge to perform like the domain expert in the domain area. Thus, validation searches for faults in the KBS behavior when it attempts to find a solution for a certain domain problem.

Usability: is an association between the KBS and the end-user. This means whether the end-user is satisfied when he/she interacts with the KBS. Therefore, it must be evaluated before installing the KBS to the end-user.

Usefulness: refers the association among the new KBS, the end-users, and the company that owns the product. The usefulness view can be noticed when the new KBS accomplishes its job. It is not possible to evaluate the new KBS if it is not functional.

2.1.5 TOOLS FOR CBR AND DM

There are different data mining and case based reasoning tools that are used for analysis and development of CBR system. Some of the tools are discussed below.

2.1.5.1 TOOLS FOR CBR

Base on accessibility, there are two types of CBR tools which are Commercial Tools and Open Source Tools.

Commercial Tools: Tools are needed in order to facilitate case collection, indexing, evaluation, adaptation, and for case library maintenance. Current commercial tools are mainly oriented towards acquisition and retrieval of cases and also simple adaptation and evaluation. The best known listed in [2] are:

- ***REMIND:*** by Cognitive Systems Inc. (USA), which is an interactive generic tool for rapid prototyping and development of CBR applications oriented to classification, prediction, and data mining tasks,
- ***recall:*** by ISoft S.A. (France) which is also a generic tool that has been applied to develop applications on fault diagnosis, bank loan analysis, teaching, risk analysis, control and supervision. An interesting aspect of ReCall is that the object oriented representation language used to represent the cases allows one to represent fuzzy knowledge.
- ***S3-Case:*** by tecInno GmbH (Germany) oriented to diagnostic problem solving and CBREXPRESS

Open Source Tools: Some commonly used free open source tools for developing CBR Application are listed and described below.

- ***myCBR:*** is one of the most popular CBR software platforms. It is a framework with certain capabilities and limitations. The most popular version of myCBR is a plug–in of open source ontology editor Protégé, but there are available some web-based versions and integration into other software. myCBR is developed by the German Research Center for AI [34].
- ***jCOLIBRI:*** is a technological evolution of COLIBRI and it is an object-oriented framework in Java which is designed for building CBR systems. It is a java-based and uses JavaBeans technology for case representation and automatically generation of user interface. This framework is developed by the GAIA artificial intelligence group in Complutense University in Madrid. The framework is built in two hierarchical levels- upper and lower. The lower level consists of library of classes (Software modules) for full 4REs CBR cycle. The upper level is "black box" graphical interface, which allows non-complicated user CBR application generation based on lower levels modules [35].
- ***Eclipse:*** is a product of Haley Enterprises. It closes relative to ART. Functionality of ART is written in LISP. Eclipse is implemented in C by NASA. In late 1980, the former chief scientist of Inference developed a new language like Eclipse. Eclipse is available for Dos,

Windows and UNIX platforms. Optimizes forward chaining using the Rete Algorithm. Can integrate data from heterogeneous databases [36].

Comparison of jCOLIBRI VS myCBR

The comparison of jCOLIBRI and myCBR from different points of view are listed and described in the table below [37].

Table 4: Comparison of jCOLIBRI and myCBR

jCOLIBRI	***myCBR***
supports full CBR cycle (retrieve, reuse, revise and retain)	supports only Retrieve and Retain
Allows retrieval from clustered and indexed case bases and submits program interfaces (connectors) to access text and XML files, as well standard and descriptive logic databases.	Does not work with external database. It stores the cases in the text file or in an XML file.
There are many CBR applications, developed on jCOLIBRI based: additional shells for distributed CBR systems, multi-agent supervisor systems for text file classification, and many CBR recommender systems.	It is valid regarding the interfaces to real-time or diagnosis systems

2.1.5.2 TOOLS FOR DM

Some tools for statistical and machine learning purpose are listed and described [38] below.

- ***RapidMiner:*** Written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. RapidMiner provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts.
- ***WEKA:*** The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms

for data analysis and predictive modeling. It is a big plus compared to RapidMiner, because users can customize it however they please. WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection.

- ***Orange:*** Python is picking up in popularity because it's simple and easy to learn yet powerful. Hence, when it comes to looking for a tool for your work and you are a Python developer, look no further than Orange, a Python-based, powerful and open source tool for both novices and experts. It also has components for machine learning, add-ons for bioinformatics and text mining.

2.2 RELATED WORK

This chapter contains three sections which are CBR, DM Techniques and Integration of CBR and DM Techniques. In each of the three sections different related research papers are studied and compared with each other and also with our research project in order to find out gap and to look the strength of their works.

2.2.1 CASE BASED REASONING

CBR as the name indicate, it is an artificial reasoning system that is based on the experts experienced cases. A new problem is solved by finding a similar past case and reusing it in the new problem situation [2]. This is because one of CBR assumption is that similar problem will have similar solution. It means that if someone got a problem and solve it using appropriate solution, after some time gone he get similar problem that he solves in the past. So now he will apply that solution that it solves his problem because the problem is similar.

According to [39], researchers proposed case representation and retrieval for statistical process control. The first focused area of the researchers is case representation. In their research indexes are chosen to represent the important aspects of the case.

As we know case indexing are used for effective similarity-based case retrieval. The researchers extract main concepts and important attributes of the issue in order to determine the indexes used for representing the cases of statistical process control. These case indexes are classified in three different groups which are functional indexes which describe the function of the case in the field

of SPC and has functional role. In functional role each case based on the defined problem goes through a path to address the problem. The second group is structural indexes represent features which describe the structural component of a case. These indexes include process features, qualitative characteristics and model features. The third index group is applied indexes which describe the application area and the date of case implementation. The second focus area of the researchers is case retrieval. The retrieved case will be the one that is most similar compared with other cases. Matching is the process of comparing two cases with each other and determining their degree of similarity (DOS). To calculate the similarity between two cases: First, corresponding features in the compared cases must be found. The values of corresponding indexes must be available in both compared cases while measuring the DOS. Otherwise, the indexes are ignored. Then, DOS between the corresponding indexes of the compared cases is computed. Finally, the obtained value is multiplied by the corresponding important factors which can be identified by the user for each index and then they add up to get overall similarity value [39].

According to [40], the author described a CBR technology for medical diagnosis and proposed architecture for CBR methodology and finally proposed specific architecture for cancer disease diagnosis. The algorithm of interpreting and assimilating a new case can be summarized in the following processes of CBR methodology architecture.

Assign Indexes where the features of the new case are assigned as indexes characterizing the event. Retrieve is the second phase where the indexes are used to retrieve a similar past case from the case memory then Modify where the old solution is altered to conform to the new situation, resulting in a proposed solution. Test is the fourth phase where the proposed solution is tried out. It either succeeds or fails. Assign and Store If the solution succeeds, then assigns indexes and stores a working solution. Then finally Explain, Repair and Test If the solution fails, and then explain the failure, repair the working solution, and test again. The explanation process identifies the source of the problem and incorporated in to the indexing rules knowledge structure to anticipate this problem in the future. The author uses the following knowledge structure to perform CBR process. The knowledge structure contains criteria in the form of rules to apply on CBR process. Indexing Rules KS (Knowledge Structure), Case Memory KS, Similarity Rules KS, Modification rules KS and Repair Rules KS.

The Author uses the CBR-based expert system prototype [40] for diagnosis of cancer diseases developed by Medical Informatics Group at Ain Shams. Patient cases are retrieved in dialogue with similarity matches using nearest neighbor techniques.

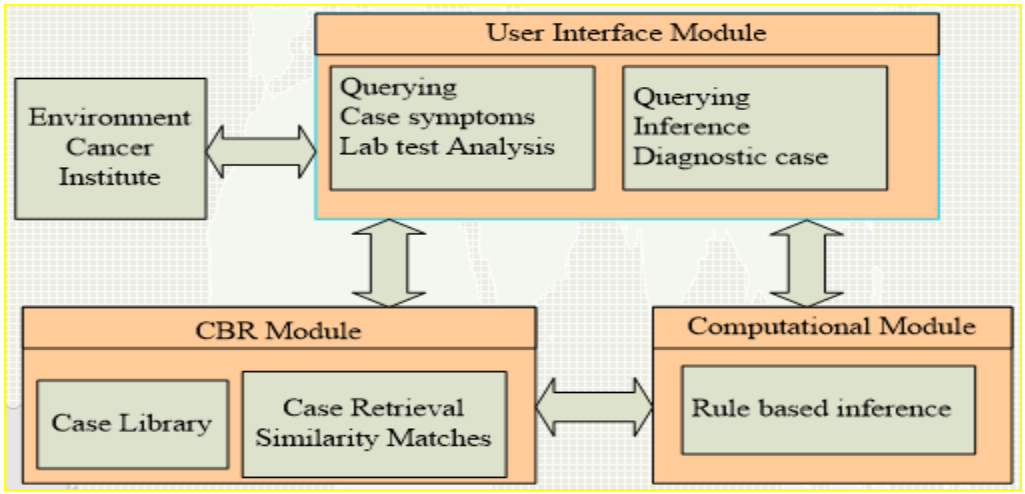

Figure 5: Architecture of CBR based Expert System for Cancer Disease Diagnosis [40]

In case based reasoning system development there are 5 processes of phases which are representation, retrieve, reuse, revision and retain. From all those phase representation and retrieve are core processes in CBR. This is because of mainly they are indicating two important parameters of system performance which are accuracy and fast response time. In case retrieval there are many algorithms and techniques are applied. One of these techniques is Euclidian Distance which is to compute the distance between the attributes of the new case and previously stored cases [41]. Each attribute are given weight by the domain expert based on their importance. The range of the distance measure of the two cases can be normalized into 0 to 1. So the Weighted Euclidian distance formula for calculating the distance between the input case and the old cases that stored in the case library.

The authors use myCBR and portge tool in order to generate results of Breast Cancer data sets from UCI Machine Learning Repository. Conversely authors Abdel-Badeeh et al [42] uses nearest neighbor retrieval algorithm to retrieve the most similar cases from the case base using a distance calculation.

Two tasks are important in enhancing the performance of case based reasoning which is Similarity measurement and case representation. Retrieval phase is the key process in CBR because it

establishes the foundation for the overall effectiveness of CBR system. It used to retrieve the most similar case from the case base to solve the user's problem by adapting similar solution. In retrieval phase the important function depend on the retrieval algorithm and similarity assessment or distance measurement.

The other task is case representation which also has influence on effectiveness of retrieval of cases. Three major types of case representation have arisen: feature vector (or propositional) cases, structured (or relational) cases, and textual (or semi-structured) cases. The feature vector or proportional cases represent a case as a vector of attribute value pairs that support K-nearest neighbor matching and instance based learning. One of the advanced approach for case representation is hierarchical case representation which store case at multiple level of detail using multiple vocabularies. When a new problem comes from the user, similar cases at appropriate level of abstraction are retrieved from the case base and the solution from these cases are combined and refined. Mainly cases contain two parts which is the problem and the solution part in some research and project they include the outcome part. The problem part describes the state of the world when the case occurred. The solution part states the derived solution to that problem and in the outcome part which will be described is the state of the world after the case occurred [28].

The new indexing method [43] used to facilitate the retrieval of relevant cases. This new index method is based on the query sphere algorithm. The researchers propose generic algorithm to search relevant cases in a discretized case base avoiding the boundary case issue. The neighborhood query problem consists in finding the relevant cases within a given distance from a given center location QC, i.e. target problem. For that purpose, the researchers adapt the spherical indexing method presented by creating an efficient domain independent indexing method. The query sphere consists of a center location and a given radius within which nearby objects must be found. Space is discretized in cubic cells. This algorithm introduces an indexing schema that gives the list of all the cells making up the query sphere for any radius and any center location. The discretization of the case base is only made on numerical features, because for nominal values it is more difficult. it demonstrates some difficulties in the situation whenever case attribute values were unevenly distributed.

2.2.2 DATA MINING TECHNIQUES

The main aim of the data mining process is to identify and extract valid novel, potentially useful information from the dossier of data and mold it into an understandable structure for future use. These understandable structure is correlations and patterns that will be considered as a new knowledge [44] and [24]. Data mining techniques have been using for different applications such as crime detection, risk evaluation, market analysis and recently for disease diagnosis.

The researchers in [45], use the CRISP-DM data mining process methodologies in order to diagnosis neonatal jaundice. As we know in the CRISP-DM there are about 5 phases which all are industry and tool neutral data mining process model that which would make the development of small as well as large data mining projects. The phases used by the researchers are Business understanding which examine different data mining tools and select the best one which is weka version 3.6, mainly because of its characteristics such as user friendly tool for health professionals. The researchers compared different studies in the literatures and expect data mining techniques to induce predictions with greater accuracy than known traditional method. In Data comprehension phase the researchers did the following: First select the hospital in which research is held at the department of obstetrics centro hospitalar Tamega e sausa N.portugal. The study includes healthy new born infants with 35 or more weeks of gestation. The collected data included mother, father sibling's information, gestational information, physical exam of the new born and clinical information of the complete stay at hospital. In Data preparation phase of CRISP-DM the following tasks were included elimination, integration, recoding and calculation of variables. In Modeling phase different classification algorithms often applied in medical datasets and implemented in weka. The algorithms applied in the medical dataset are J48 (implementation of the C4.5 algorithm for generating pruned or unpruned decision tree), Simple CART (a decision tree learner implementing minimal cost complexity pruning), Naïve Bayes, Multilayer perception, Sequential minimal optimization, Simple Logistic and other similar methods were also used but without better result and therefor are not reported in there study and in their research they consider the last phase as Evaluation. The tests were performed using internal cross validation 10 folds used to determine how the quality of a learning algorithm will be affected in a separate set of data. All classification algorithms were tested for different subsets of variables and compared in terms of

accuracy, sensitivity and specificity. High accuracy was obtained bayes net algorithm (Acc=0.74) followed by Naïve Bayes and simple logistic (Acc=0.72)

According to [41], which used different data mining algorithms such as Naïve Bayes, SMO, Bagging, and Neural Network were applied on Z-Alizadeh Sani dataset for diagnosis of coronary artery disease. A new feature creation algorithm is proposed to derive three new features from existing ones and those features are used to recognize whether three major coronary arteries, Left Anterior Descending, Left Circumflex, or Right Coronary Artery is blocked. The authors use Association rule mining to obtain rules that determine the label of the patients (CAD or Normal). for experimental purpose the authors use RapidMiner tool which is an environment for machine learning, data mining, text mining and business analytics. As the experimental result show they received highest accuracy similarly [46] use Association rule mining specifically Apriori and Frequent pattern-Growth algorithm to predict diabetes disease earlier on the other hand [47] use predictive data mining algorithms such as decision tree and navie bayes algorithms to diagnosis heart diseases. For experimental purpose they use a tool called weka. In predictive data mining the researchers use five parameters which are sensitivity, specificity, precision & recall, accuracy and confusion matrix to validate the result obtain. As the result shows navie bayes has a little bit more accurate than decision tree algorithm. Similarly, [48] authors construct predictive model to predict pressure ulcers and select critical attribute for risk factors. The authors use four data mining techniques, namely, Mahalanobis Taguchi System, Support Vector Machines, decision tree, and logistic regression to select those critical attributes which leads to predict the incidence of pressure ulcers. The authors use four parameters to measure the performance of the used techniques. These are sensitivity, specificity, F1 and g-means. The results show that data mining techniques obtain good results in predicting the incidence of pressure ulcer.

The researchers in [49] briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural network to massive volume of health care data. As we know data mining and knowledge discovery from database (KDD) are different but often used interchangeably and the researchers shows how they differ in their work practically. KDD is the process of turning the low-level data into high-level knowledge but Data mining is an important step in the KDD process which applies intelligent and statistical techniques to extract patterns potentially useful. The Knowledge Discovery in Databases

process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps: Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge representation. The patterns that can be discovered depend upon the data mining tasks applied. As we know generally, there are two types of data mining tasks which are descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data. The researchers use some of the data mining key steps which are Problem definition, Data exploration, Data preparation, Modeling, Evaluation and Deployment. For data preprocessing and effective decision making one dependency augmented Naïve bayes classifier and naïve creedal classifier are used. ODANB are characterized by all attributes have certain influence on the class and the conditional dependency assumptions is relaxed. Before running the ODAND algorithm the following data preprocessing techniques are applied replace missing values, discrimination etc. The researchers use one of the comparison criteria which is Accuracy of prediction (measures defined from the confusion matrix outputs). The table below summaries the benchmarked algorithms accuracy for each dataset considers. For Diagnosis of Lung Cancer Disease Naïve Bayes (NB) observes better results than (ODAND) for all datasets as shown in the table below.

Table 5: Comparison Of Accuracy for NB and ODANB Classifier for Different Dataset [49]

DATASETS	***ODAND***	***NB***
LUNG CANCER-C	80.46	84.14
LUNG CANCER-H	79.66	84.05
LUNG CANCER-STATLOG	80.00	83.70

2.2.3 INTEGRATION OF CBR AND DATA MINING TECHNIQUES

There are many researches that have been doing on hybrid system. Particularly CBR system are hybrid with different artificial intelligent approaches such as CBR with rule base reasoning, CBR with model based reasoning, CBR with information retrieval and soft computing methods (i.e. fuzzy methods, neural networks, genetic algorithms) [50] and [51]. The reason for system hybridization is to get the advantage of both of the system and also to increase the performance of the hybrid system.

Integrating case based reasoning and rule based reasoning are common hybrid system in the artificial intelligence development approach [52]. This is because a single technique may not give optimal solution and as test result shows solutions generated by the proposed hybrid model are better than those generated by using a single technique [53], [54], [55].

In medical application it is common to use hybrid of AI paradigm with CBR but to integrate CBR and data mining techniques are rare. Mapping data mining functionalities to case based reasoning tasks and steps such as case mining, memory organization, case base reduction, generalized case mining, indexing and weight mining were done by Isabelle Bichindaritz [60]. INRECAtree which is a hybrid between a decision tree and a K-D tree. This method allows both similarity based retrieval and decision tree retrieval is incremental and speedup the retrieval. As we know data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

A model of a chronic diseases prognosis and diagnosis system integrating data mining and case-based reasoning was proposed by M.-J. Huang et al [56]. they build up Algorithm of Two-phase Knowledge Integration (Tp-KI). The first one is a knowledge creating phase which use data mining techniques, the decision tree induction algorithm and the case association to generate meaningful rules from health examination data. The second phase of knowledge inferring is based on the phases of case based reasoning. The system is verified from different perspectives and one of it is to check case retrieval accuracy and the test result is from 20 cases the similarity between the stored case and the new case is between 0.8 and 1. So it has got very satisfied result. They use data mining mainly for the purpose knowledge discovery but they can use algorithms from data mining in different phase including the retaining phase for case to be learnt by the system and to be stored.

Similarly, [57] use data mining for discovering hidden patterns in past pathology requesting data and case based reasoning methodologies is to retrieve and enable the use of this knowledge through CBR for the purposes of decision support. A bit similar with the above authors in [58] use data mining techniques which is decision tree algorithms of C5.0 and case based reasoning for retrieval purpose using KNN algorithm. To make the inference engine they use C5.0 algorithms and CBR paradigm in order to predict retinopathy.

CHAPTER THREE

METHODOLOGY

3.1 RESEARCH DESIGN

The study is all about enhancing the effectiveness (in terms of retrieval performance and increasing the accuracy of the retrieved case solution) by using hybrid techniques and paradigms, which are case based reasoning and data mining techniques. In this study, we select the application domain to be pneumonia disease diagnosis. The necessity of the study is unquestionable because the problem spread worldwide and pneumonia is the most killing disease especially in Africa. There is shortage of doctors in Africa especially in Ethiopia. The ratio of doctors to patient in Ethiopia is high. Threrefore, if we have strong disease diagnosis system that will support the patient and medical practitioners, the problem will be reduced to some extent.

The dataset was collected from Adama hospital and medical college, different standard guidelines and general practitioners. The data collected from patient medical record cards. As we specified the domain area, the data is on pneumonia disease diagnosis. The type of dataset collected from medical record cards directly from the hospital; So that our data source is primary. The type of dataset collet was Quantitative. As we know the aim of Quantitative research is determining the relationship between quantitative variables or compare groups. So that one of our aim is to find out association between symptoms, diagnosis result and medical prescription.

The techniques of data collection method we used are not the commonly used method like observation, interview, questionnaires, etc. As we explain earlier, the data is collected from medical record cards in the hospital. Threrfor, we use a tabular form to collect data and such kinds of data collection techniques are called Document Review.

For data preprocessing purpose we use weka data mining tool and problems like missing values, outliers are solved using this tool. Our research design is concerned with predictive and diagnostic research design; so that we make survey to identify the best data mining algorithm for association rule mining and classification techniques. The algorithms applied on the pneumonia dataset and we also develop a CBR application using jcolibri framework.

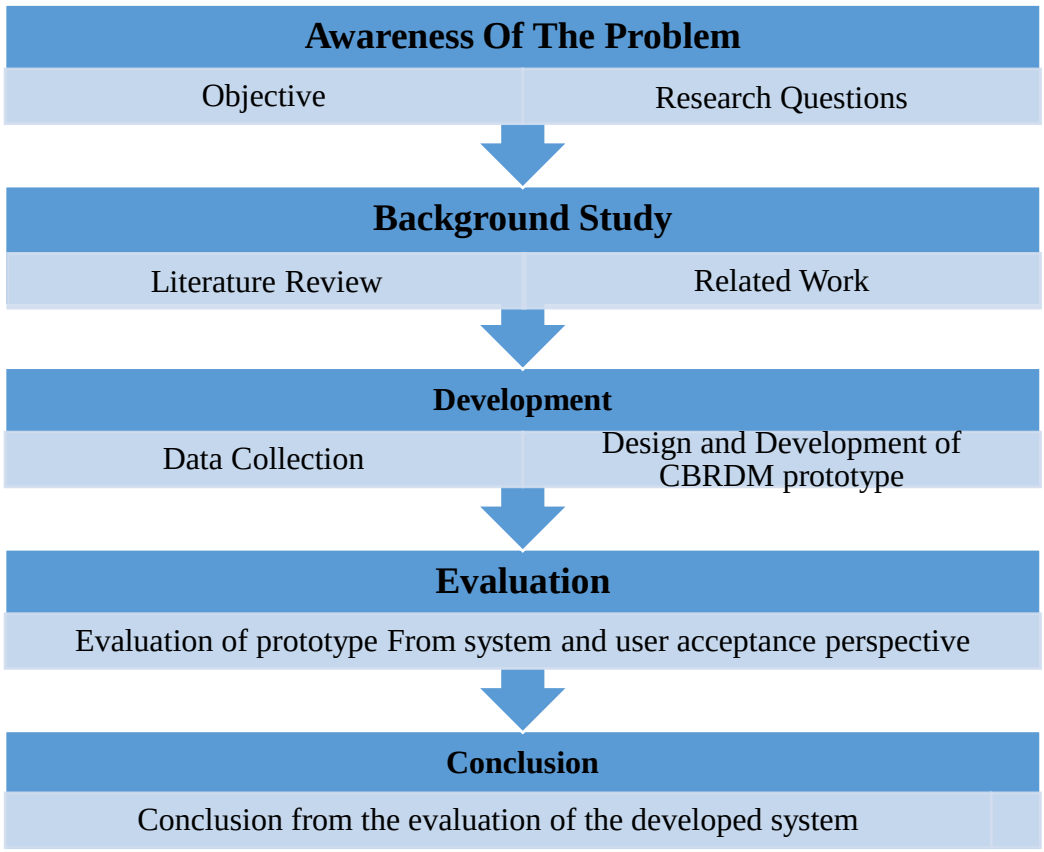

Figure 6: Milestone Phases and Tasks in the Thesis

3.2 PROCESS MODEL

Several software development approaches have been used since the origin of information technology, in two main categories. Typically, an approach or a combination of approaches is chosen by management or a development team [28].

For research purpose the recommended types of software development process model are prototyping and open source software development technique. Because since research is an activity where you either don't have a known end result, or you have observed something and are working to determine the underlying mechanism that produces your observation. It means that we don't have predefined out come and don't need to show the complete functionalities of the system [29]. We prefer to use an approach which is Evolutionary prototyping process model.

Evolutionary Prototyping: This type of prototyping is based on the idea of developing an initial implementation, exposing it to user comment and refining it through repeated stages until an adequate system has been developed. The stages are:

I ***Requirements definition (initial specification):*** a stage of thorough analysis is used to create an initial specification for the software. The system must have a query form that the user selects and write the symptom and sign what he/she fills. The prototype system must have the similarity configure dialogs Finally, Result dialog box to display the retrieved result to the users

II ***Prototype construction:*** prototype is built in a quality manner, including design, documentation, and thorough verification. In prototype construction we use open source frameworks which is jcolibri. Importing the two framework into eclipse IDE and adopting based on our design of prototype system. In more detail of how the prototype system constructed are found in implementation section.

III ***Evaluation (check with the user):*** during evaluation, problems in the developer's perception of the customer requirements are uncovered. The prototypes are the communication medium that enables the developer and customer to communicate with each other. Evaluating the prototype system with some sample users, to identify in which part of the prototype system they are satisfied and unsatisfied. This lead us to improve the prototype system until the user are satisfied.

IV ***Iteration (refine the prototype):*** evaluation is carried out repeatedly until the prototype meets the objectives. The specification is updated with every iteration.

3.3 TECHNOLOGY USED

Within the field of computer-science in general there is usually countless of open source or commercially available software packages and applications that help designers and developers in doing their job better and more efficient. Within artificial intelligence and machine learning, there really isn't such an abundance, and most systems developed in research is made from scratch [1]. Some of the technologies used in our research are discussed below.

jCOLIBRI: is a Java framework for developing CBR systems that allows users to reuse software code by providing a substantial library to both make the development more efficient, but also ensure consistency over time. In our development of a prototype CBR application we used the jColibri framework as a starting platform and extended it with classes to suit our needs [35].

Eclipse: is an Integrated Development Environment (IDE) for Java, an object oriented programming language. Eclipse has over the years become almost the defacto IDE to use for Java development, and offers a stable environment with many handy features to simplify the development process [36].

Hibernate: is an Object Relational Mapping (ORM) middleware that handles the mappings between objects in a programming language such as Java and the underlying storage, a simple database in our case.

Weka: is an open source data mining tool developed at the University of Waikato in New Zealand. The abbreviation stands for Waikato Environment for Knowledge Analysis [61]. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes.

Navicat: It is not only a powerful, sophisticated, and easy-to-use all-in-one database administration tool with a GUI, but also a very useful aide for developers who work on database-driven applications. It allows to import/export data, back up, or transfer an entire database to another server and design queries in a GUI with point-and-click and drag-and-drop features [52].

JavaFX and FXML: JavaFX is an open source Java-based framework for developing rich client applications. JavaFX is also seen as the successor of Swing in the arena of graphical user interface (GUI) development technology in Java platform. The JavaFX library is available as a public Java application programming interface (API) [63].

FXML maps directly to Java, most JavaFX classes can be used as elements, and most Bean properties can be used as attributes. It is particularly useful for user interfaces that have large, complex scene graphs, forms, data entry, or complex animation. FXML is also well-suited for defining static layouts such as forms, controls, and tables. In addition, you can use FXML to construct dynamic layouts by including scripts [64].

From a Model View Controller (MVC) perspective, the FXML file that contains the description of the user interface is the view. The controller is a Java class, optionally implementing the Initializable class, which is declared as the controller for the FXML file. The model consists of domain objects, defined on the Java side, that you connect to the view through the controller [64].

3.4 DATA SOURCES

Medical records are all types of data which obtain from the hospital environment and related to patient diagnosis and treatment. Medical records include clinical information obtained by case history, physical examination, data acquired through various diagnostic procedure, information related to various therapeutic interventions and data which are of administrative and of financial importance such as insurance, costs of medical treatment and costs of hospitalization.

Clinical data set is not easily accessed since the data are so sensitive and needs confidentiality because they contain patient diagnosis information. We got an ethical letter from ASTU and went to different hospital, finally to Adama hospital and medical college and showing the cooperative letter and asked what type of data we want from the hospital. Finally, they accepted the letter but there was another challenge which was the data set wasn’t in the soft copy digital format rather than manually in paper. Another important source for the dataset is standard medical guideline and case studies on the pneumonia diseases. The source is got from [65], [66].

3.5 DATA COLLECTION TECHNIQUES

Even if there are different techniques to collect data such as Observing, Interviewing (face-to-face), Administering written questionnaires, Focus group discussions etc. but in this thesis the suitable and preferable way is to get dataset from the clinical database. Because we need many number of cases to be represent in the case base and to generate rules so that rules to be represent in the rule base. But there is no clinical database that contains all patient information related with pneumonia. The other optional way to collect data was to get data from clinical patient medical record cards. The techniques is called Using available information (Document review).

Document review is a large amount of data that has already been collected by others, although it may not necessarily have been analyzed or published. Locating these sources and retrieving the information is a good starting point in any data collection effort. Tool to collect data are Checklist and data compilation forms. This types of data collection techniques are inexpensive, because data is already there. It permits examination of trends over the past. The disadvantage of such data collection techniques are Data is not always easily accessible, Ethnical issues concerning confidentiality may arise and Information may be imprecise or incomplete [67]. The process is very tedious and also needs professional person on the area of pneumonia disease. The steps in the process of data collection are mentioned below:

- ***Identifying focus group to collect the data:*** as indicated earlier there are two focus areas to collect data which are medical ward for adult and children (under 5) and young department. Why the two focus areas are selected? It is because of the disease pneumonia affect mainly children and the data related with pneumonia are mainly found in the two department.
- ***Selecting card numbers from all types of disease registration books which are hospital medical registration(HMRS):*** Card numbers with related types of pneumonia are selected and registered. 230 card numbers which indicate pneumonia disease were identified from HMRS
- ***Contract with professional health personal:*** to write a case (all the attributes such as symptoms & signs, related pneumonia types and the treatment (prescription)) in the card rooms.

- ***Preparing structured form:*** to register the selected case which contains medical registration number (MRN), Age, Sex, Symptom/Sign, Diagnosis result and prescription (Rx).

Table 6: Form to Collect Data

Case No	***MRNS***	***Age***	***Sex***	***Symptom & Sign***	***Diagnosis***	***Prescription***
1	607316	18	F	Cough, SOB, Fatigability	Streptococcus Pneumonia	Amoxicillin625mg, doxycycline100mg PoBiD #7

- ***Registering all the cases from the card rooms:*** There are thousands of patient registration cards in the card rooms. Some cards were lost, some others were damaged and some others were completely unrelated with pneumonia disease. From 330 identified card numbers only 200 cards correctly contain pneumonia disease type and the necessary data.

3.6 SAMPLING TECHNIQUES

We usually would have one or more specific predefined groups we are seeking. In our case when we go to hospital we have many purposive words in our mind, pneumonia disease, adult, children and their gender and so on. In this research we use quota sampling which we select people non-randomly according to some fixed quota. In this research, we prefer to use non-proportional quota sampling which is a bit less restrictive. In this method, we specify the minimum number of sampled units we want in each category. The sample category includes Gender (Male, Female), Age (all age groups). Here, we're not concerned with having numbers that match the proportions in the population. Instead, we simply want to have enough to assure that we were able to represent and talk about even small groups in the population.

3.7 DEVELOPMENT OF THE PROTOTYPE

The prototype system is the integration of CRISP-DM sequential process followed by jCOLIBRI framework for CBR application.

3.7.1 DATA MINING PROCESS USING CRISP-DM

CRISP-DM are explain in the next chapter. The details of how we use CRISP-DM process model in this research will be seen in chapter 4.

3.7.2 CASE BASED REASONING USING JCOLIBRI2

jCOLIBRI2 Framework has two main layers the white box bottom layer (java developers) and the black box top layer (CBR designers). We use the white box bottom layer since the layer has advantages and explain in [35] below.

White-box framework: reused mostly by sub classing. The new design of jCOLIBRI2 attempts to remodel the architecture into a clear white-box system oriented to programmers. It takes advantage of the new possibilities offered by the newest versions of the Java language. The most important change is the representation of cases as Java Beans. The persistence of cases is managed by the Hibernate package. Hibernate is a Java Data Objects (JDO) implementation, so it can automatically store Java Beans in a relational data base, using one or more tables. Java Beans and Hibernate are core technologies in the Java 2 Enterprise Edition platform that is oriented to business applications.

Black-box framework: reused through parametrization has an associated visual builder that will generate the application's code. black-box with builder layer that is oriented to designers. Regarding the top layer, it is oriented to designers and includes several tools with graphical interfaces. This layer is the black-box version of the framework, helping users to develop complete CBR applications guiding the configuration process.

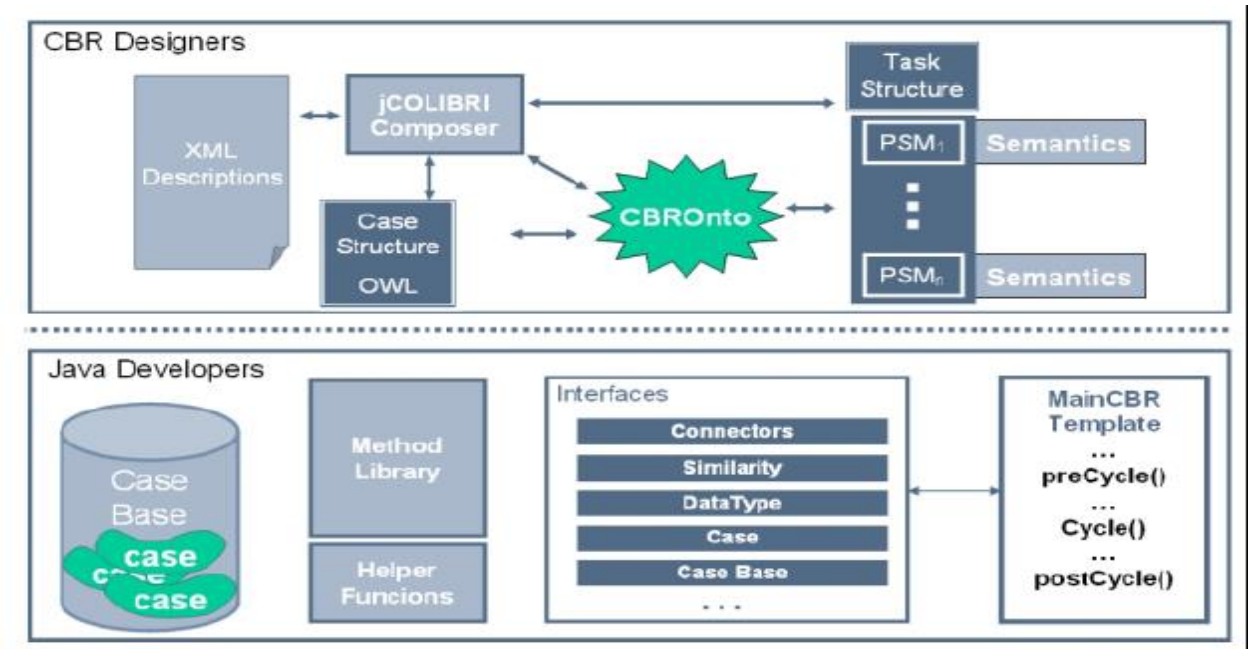

Figure 7: Two Layers Architecture of jCOLIBRI2 [35]

3.8 PROGRAMMING LANGUAGE USED AND JUSTIFICATION

Programming language used depend on the software framework or tools. The software framework used are JCOLIBRI2 case based reasoning framework and WEKA suite of data mining algorithms which are an open source software and is fully dependent on java programming language.

3.9 TEST PROCEDURE

In this research two methods of evaluating the prototype are used. The first method is using the system performance metrics and the second evaluation method is user acceptance test. The first method use different metrics such as accuracy, retrieval performance (Recall and Precisions) and the second evaluation method use standard user acceptance test metrics such as Completeness, Flexibility, Usability, Functionality, Ease of installation, Performance, Integrity and so. From this metrics user acceptance evaluation questions are prepared and given to domain expert to evaluate the prototype system.

CHAPTER FOUR

PROPOSED INTEGRATION ARCHTECURE OF DATA MINING TECHNIQUE AND CBR

The system is designed to combine the two reasoning paradigms of CBR and DM techniques in order to give as accurate result as possible. The CBR part of the system stores cases in a case-base with saved symptom description and their matching diagnosis result and doctor's prescription. It is unusual to use DM techniques as reasoning techniques. DM has different techniques and from all those techniques, we select association rule mining and classification in order to generate general rules about pneumonia symptom description and diagnosis result and doctor's prescription.

The basic idea was to have the DM work as a type of RBR where the specific cases was looked up and guided by generalized knowledge. Basically what would happen was that a new case would be presented the system which would do the regular CBR attribute matching based on some given similarity metric. Correspondingly the general rules which are most similar with the specific case are retrieved in the system. The Architecture of integration of case based reasoning and Data Mining Techniques are shown in figure 8 below. In the DM techniques we, use the CRISP-DM process cycle and CBR cycle part we use well known 4R CBR process cycle by A. Aamodt and E. Plaza (1994).

In the figure 8 below, according to CRISP-DM the Business understanding (Phase I) and data understanding (Phase II) are not seen in the integrated Architecture because as we know they are not use any tools, and the two phases are about understanding the overall data mining projects including the dataset. In the data preprocessing and preparation phase we just try to keep consistency of the dataset by calculating any missing values, outliers and other attributes of noise data. Then after that the dataset move into two directions. The first one is to the case base to be applied on CBR and the other direction is to move down to the fourth phase (Modeling) of CRISP-DM to generate rules using more satisfying techniques and algorithms.

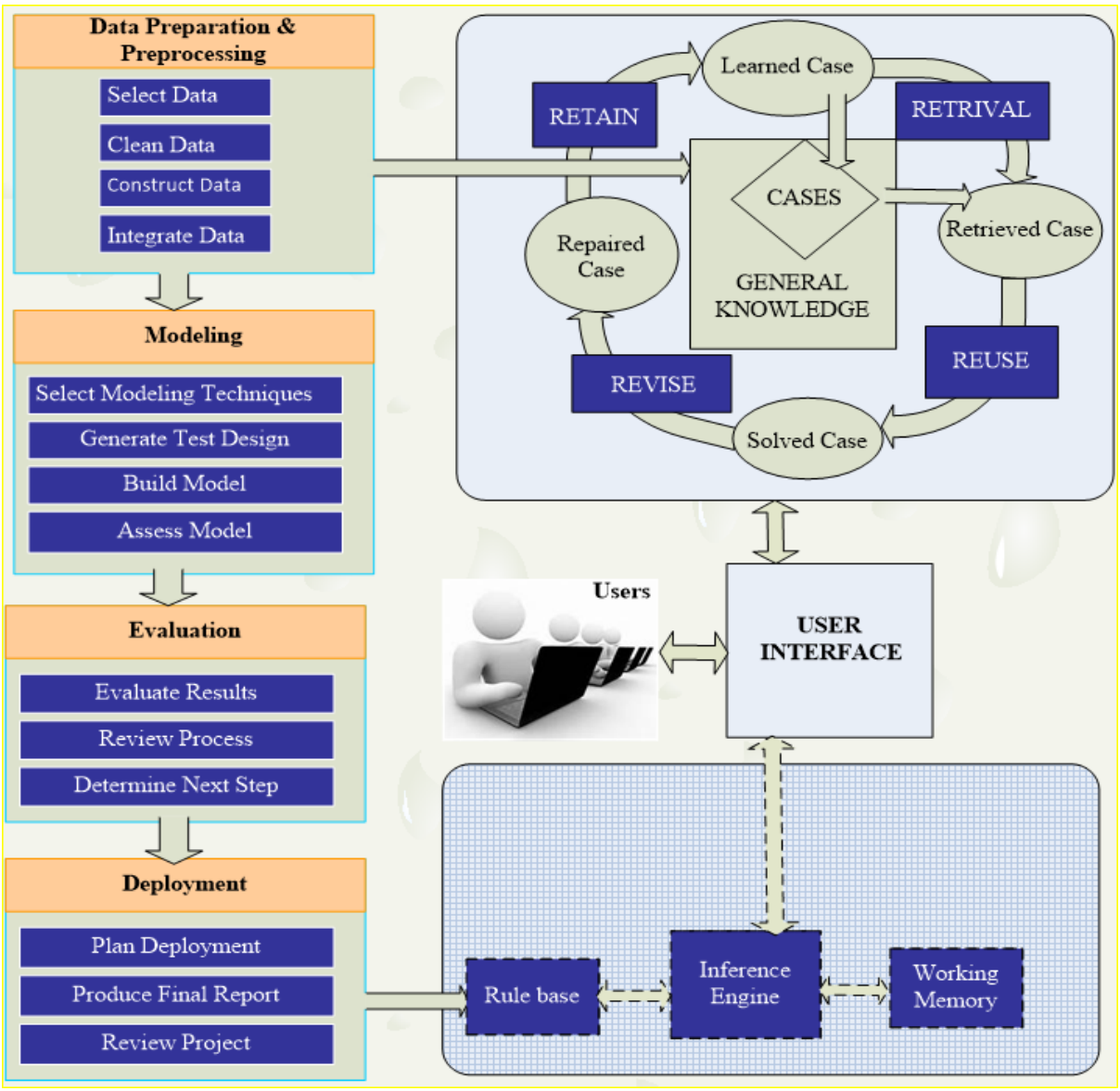

Figure 8: Integration Architecture of DM Technique and CBR

The architecture of the system is to integrate data mining process methodology with case based reasoning paradigm. In this CBRDM architecture DM and CBR are fully implemented but the RBR part are shown only on the architecture. The rules are produced using the CRISP-DM process model but those rules are not implemented on RBR part of the prototype system.

According to [10], there are three main approach to integrate CBR with other techniques and paradigms. In this research, what we realize is the integration of data preparation and preprocessing step of CRISP-DM process model with CBR is loosely coupled sequential

integration. It is because the output of prepared and preprocessed dataset will be the input case for CBR system.

The other loosely coupled sequential integration in the architecture is the rules produced by CRISP-DM model are an input general knowledge to RBR. The architecture has a common interface for receiving reasoning results from CBR and RBR. The two reasoning paradigms operate correspondingly and provide solution and conclusion that support each other. Such kinds of integration are cooperation Coprocessing integration. In our research the general rules support the specific case results from CBR.

4.1 CRISP-DM IN CASE OF CBRDM

CRISP-DM is an industry- and tool-neutral data mining process model that which would apply on small as well as large data mining projects faster, cheaper, more reliable and more manageable. In the CRISP-DM data mining process cycle there are 6 steps which are some of them are iterative. The steps are:

4.1.1 BUSINESS UNDERSTANDING

The selected data mining tool is weka version 3.7 because of its better characteristics such as user friendly tool and it contains many data mining techniques and algorithms. The objectives of CRISP-DM in this research perspective are: To clean and find out relevant rules from the given medical data set using data mining techniques. Adama hospital medical college has main objective which is protecting the society from disease and gives health service including diagnosis and medication or medical treatments. The hospital is also health science collage and the dataset will use for medical practitioners to studies different cases.

4.1.2 DATA UNDERSTANDING

The data was collected from both pediatrician and medical ward departments of medical ward A and medical ward B from Adama Hospital medical college. The other dataset source was standard medical guidelines and general practitioners manuals. The total dataset has 1006 records (instances/tuples) and it have 22 attributes. The attributes have 4 different classes which are patient profiles, sign & symptoms, diagnosis result and medication (treatments). The attributes are MRNS, sex, age, cough type, fever Type, loss of appetite, unable to feeding, sweating type, Shortness of

breath, fast breath, chest pain, chest in drawing, weight loss, chills rigors, grunting, vomiting, AFB= negative, nasal flavoring, CBC, CXR, pneumonia Type, Prescription.

Table 7: Detail Description of the Dataset Attributes ([68])

S.N	Attributes	Meaning/Description	Values	Data Types
1	CaseId	Unique numbers which are given for patient to be identified	Numbers	Integer
2	Age	The Age of patients		
3	Gender	The sex of the patients	Male	nominal
			Female	
4	Cough type	A cough is a sudden and often repetitively occurring, protective reflex, which helps to clear the large breathing passages from fluids, irritants, foreign particles and microbes. A productive **cough** helps clear mucus (sputum) and foreign material from the airways. Dry **coughs** are usually caused by irritation from cigarette smoke, environmental irritants, allergies, postnasal drip, or asthma. No Cough refers to there is no observed symptom or sign of cough.	Cough	Nominal
			Productive Cough	
			Dry Cough	
			No Cough	
5	Fever Type	A fever occurs when the body temperature rises above normal. The normal body temperature is between 36 and 37°C, but this can vary from person to person and from hour to hour. Temperatures between 37.5°C and 38.2°C mark a low-grade fever. A high-grade fever is present when the oral temperature is above 38.2°C.	Mild Fever	nominal
			High Grade Fever	
			Low Grade Fever	

6	Appetite for Food	A feeling of craving for food	Good/ Poor	Binary
7	Sweating	Night sweats caused by a medical condition or infection can be described as "severe hot flashes occurring at night that can drench sleepwear and sheets, which are not related to the environment.	Present/ Absent	nominal
8	Shortness of breath	When there is inflammation of the alveoli in pneumonia, their gas-exchanging functionality is naturally compromised. The alveoli are unable to extract sufficient oxygen from the air, which thus produces the feeling of breathlessness.	Present/ Absent	Binary
9	Chest pain	A very common symptom of pneumonia is chest pain. Either it usually manifests as a sharp, stabbing pain or a dull ache, of which occurs when the patient is inhaling or exhaling.	Present/ Absent	Binary
10	ChestIndrawing		Present/ Absent	Binary
11	NauseaVomitingDiarrhea	Discharge of liquid from the body. ***Nausea***: The state that precedes vomiting, Disgust so strong it makes you feel sick. ***Vomiting:*** the reflex act of ejecting the contents of the stomach through the mouth. ***Diarrhea***: Frequent and watery bowel movements; can be a symptom of infection or food poisoning or colitis or a gastrointestinal tumor.	Nausea Vomiting Diarrhea Absent	Binary
12	Complete Blood Count	A complete blood count (CBC) is commonly done before and regularly during treatment. The CBC shows the	Positive Negative	Binary

		numbers of white blood cells, red blood cells, and platelets in the blood	Not Checked	
13	Chest-X-Ray	To look for inflammation in your lungs. A chest x ray is the best test for diagnosing pneumonia. However, this test will not tell your doctor what kind of germ is causing the pneumonia.	Positive Negative Not Checked	Binary
14	PneumoniaType	Community-acquired pneumonia (CAP) occurs outside of hospitals and other health care settings. Most people get CAP by breathing_in germs (especially while sleeping) that live in the mouth, nose, or throat. Streptococcus pneumoniae: usually caused by inhaled bacteria called Streptococcus Atypical pneumonia: It does not cause symptoms that require bed rest. It might just feel like a common cold and can go unnoticed as pneumonia. Bronchial pneumonia: Acute inflammation of the walls of the smaller bronchial tubes, with irregular areas of consolidation due to spread of the inflammation into the alveolar ducts of the lungs. Pneumonia: not specifically known pneumonia.	CAP S.CAP Pneumonia S.Pneumonia Atypical Pneumonia Broncho Pneumonia	nominal
15	Prescription	Medical treatment that ordered by a doctors	Various	

4.1.3 DATA PREPARATION AND PREPROCESSING

The data is collected manually, it means that the data is not collected from data base rather than it gathered and copied from paper based patient registration card. Because of this most of data preparation processes are covered at initial data collection stage. In the data preparation process different operations are performed such as which data features are relevant to meet the research objectives? There are different data features in a medical record such as symptom, sign, patient profile, date of entrance and date of exit for a patient, diagnosis Result, prescription and treatment, daily follow up of patient and so on data features are accumulated in a single medical record numbers. Out of this data features the relevant with regard to the business objective are symptom, sign, patient profile, Pneumonia Types, Prescription.

Selecting Data: Reducing the dimension of the dataset by performing feature selection. Some of the data was not pertinent to the data mining exercise, and was ignored. Of the variables given in Table 1 above, MRNS was ignored as having no data mining value. It was, however, used during preprocessing and post-processing to aid in data selection and gaining better understanding of rules generated. Some of the data has very low weight and not commonly seen in pneumonia diagnosis. So we removed those feature or attributes such as Weight loss, Chills rigors, Grunting, AFB= negative.

Data Cleaning: The data which include tasks such as removing duplicates, removing inconsistencies, supplying missing values, etc. Cleaning involves identification of missing, inconsistent, or mistaken values. Tools used in this process step included graphical tools to provide a picture of distributions, and statistics such as maxima, minima, mean values, and skew. Some entries were clearly invalid, caused by either human error or the evolution of the problem reporting system. For instance, over time, input for the Class attribute changed from SW-bug to sw-bug. Those errors that were correctable were corrected. If all errors detected for a report were not corrected, that report was discarded from the study.

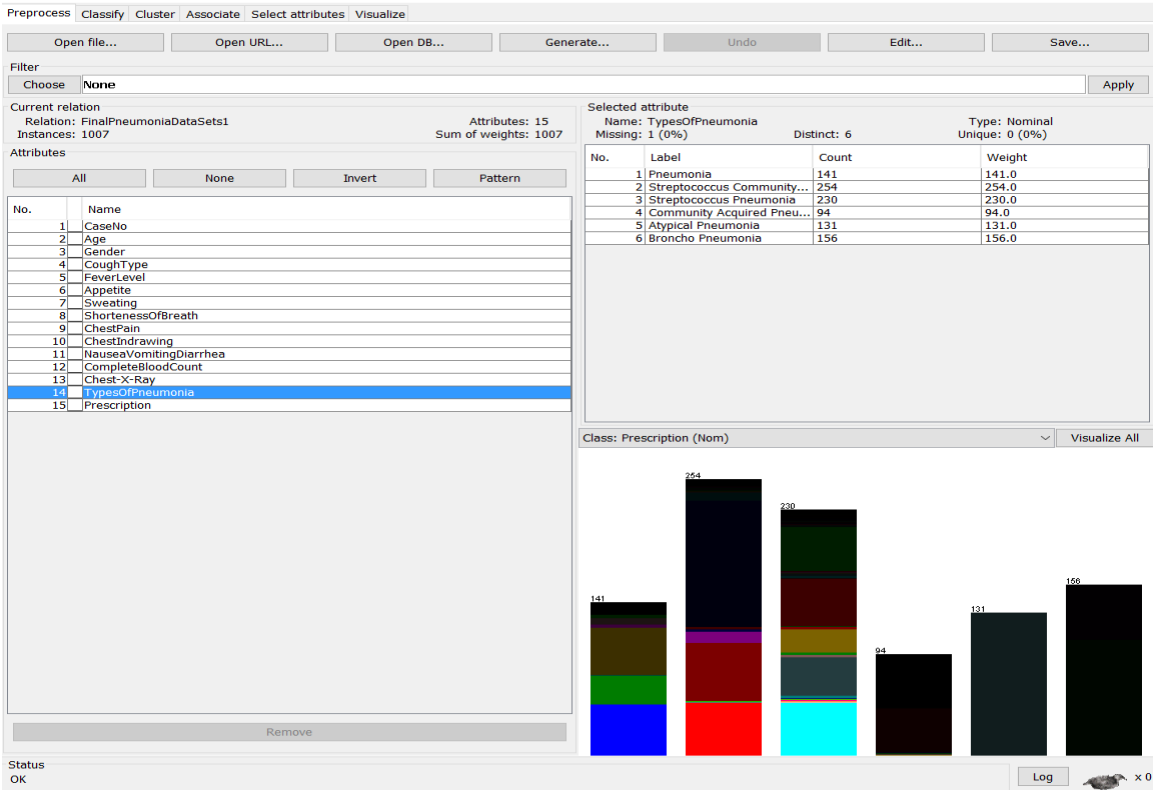

Figure 9: Snapshot of 15 Attributes of Patient from Weka DM Tool

The above figure 9, shows that all attributes of patient and especially the bar graph shows 6 types of Diagnosis result of pneumonia disease (Community Acquired Pneumonia, Streptococcus Community Acquired Pneumonia, Pneumonia, Streptococcus Pneumonia, Atypical Pneumonia, Broncho Pneumonia) versus Prescription which has many types doctor's prescription of medicine to patient. The graph shows outliers, distinct, unique and missing values of the selected attribute diagnosis result. The table 8 below is created in this manner by selecting all attributes step by step and looking for values of outliers, distinct, unique and missing values.

Table 8: Different Values Before Preprocess of Attributes From the Weka

Attributes	***Outliers***	***Distinct***	***Unique***	***Missing Values***
Case No		1006	1006	1(0%)
Age		76	0(0%)	1(0%)
Gender		2	0(0%)	1(0%)

Cough Type		4	0(0%)	1(0%)
FeverLevel		4	0(0%)	1(0%)
Appetite		2	0(0%)	1(0%)
Sweating		2	0(0%)	1(0%)
ShortnessofBreath		2	0(0%)	1(0%)
ChestPain		2	0(0%)	1(0%)
ChestIndrawing		2	0(0%)	31(3%)
NauseaVomitingDiarrhea		4	0(0%)	43(4%)
CompleteBloodCount		3	0(0%)	38(4%)
Chest-X-Ray		3	0(0%)	1(0%)
TypesOfPneumonia		6	0(0%)	1(0%)
Prescription		76	42(4%)	1(0%)

As we see in the table 8 above; there is no Outliers value, but there is missing values which needs to be filtered. In this thesis we replace all missing values for nominal attribute in a dataset with the modes calculation. pneumonia dataset for all attributes and so we did not try to handle any missing data and Outliers.

Construct Data: There are two ways to construct a given dataset. The first one is Derived attributes These are new attributes which are comprised of one or greater existing attributes in the same record, for example you might use the variables of length and width to calculate a new variable of area. But in our case we are not get any derived attributes from the existing attributes. The second way is Generated records - Here you describe the creation of any completely new records. For example you might need to create records for customers who made no purchase during the past year. There was no reason to have such records in the raw data, but for modelling purposes it might make sense to explicitly represent the fact that particular customers made zero purchases. Again we do not need to have a generated records in the pneumonia dataset.

Integrated Data: some of pneumonia symptoms are called in different name but have the same meaning. For example, “loss of appetite” and “unable to feed”, “Fast Breath” and “Shortness Of Breath” have similar meaning so that we selected one of the two. What we select is Appetite and to have Good and Poor values which is a common word, similarly we selected ShortnessOf Breath which is commonly used symptom and assigned values are Present and Absent [5].

4.1.4 MODELING

There are different data mining techniques that could be applied in the given data set. But the question is which data mining techniques can satisfy the business objectives? As mentioned in the first phase of CRISP-DM process the main objective is to find out relevant rules from the available pneumonia datasets. The following sub section are described and applied on the modeling of CRISP-DM.

Select Modeling Techniques: As the first step in modeling, select the actual modeling technique to be used. If a tool was selected in business understanding, this task refers to selecting the specific modeling technique, e.g., building decision trees or generating a neural network. Some of the Data Mining techniques are classification, clustering, prediction, association rule mining and so on. On some of the data mining techniques, we apply analysis using tools, on some of the others systematic survey are applied.

- ***Classification:*** divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as “high risk” or “low risk” patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, “high” or “low” risk patient may be considered while the multiclass approach has more than two targets for example, “high”, “medium” and “low” risk patient. Data set is partitioned as training and testing dataset. Using training dataset, we trained the classifier. Correctness of the classifier could be tested using test dataset [69].
- ***Clustering:*** is an unsupervised learning method that is different from classification. Clustering is unlike to classification since it has no predefined classes. In clustering large database are separated into the form of small different subgroups or clusters. Clustering

partitioned the data points based on the similarity measure [69]. Clustering approach is used to identify similarities between data points. Each data points within the same cluster are having greater similarity as compare to the data points belongs to other cluster [69]. Given n samples without class labels, It is sometimes important to find a meaningful partition of the n samples into c subsets or groups. Each of the c subsets can then be considered a class by themselves. That is, we are discovering the c classes that the n samples can be meaningfully categorized into. The number c may be itself given or discovered. This task is called clustering [69].

- ***Prediction:*** is very similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for unseen objects [28].
- ***Association:*** is one of the most vital approach of data mining that is used to find out the frequent patterns, interesting relationships among a set of data items in the data repository [69]. An association rule is pair that we write as $X \Longrightarrow Y$, where X and Y are two itemsets and $X \cap Y = \emptyset$. The itemset X → antecedent of the rule. The itemset Y → consequent of the rule.

There are different association rule mining algorithm to extract useful and novel rules from the available dataset some of them are described below. The algorithms that use association rules are divided into two stages, the first is to find the frequent sets and the second is to use these frequent sets to generate the association rules. The two commonly used association rule mining algorithms in weka are Apriori and FP Growth. We describe the two algorithms below.

I ***Apriori:*** generates less candidate itemset for testing in every database pass. The search for association rules is conducted by two parameters: support and confidence. Apriori yield an association rule if its support and confidence values are above user defined threshold values. The output is ordered by confidence. If various rules have the same confidence, then they are governed by support. This algorithm uses a breadth first search approach. Apriori implementation uses a data structure that directly represents a prefix tree. The tree grows top-down level by level, removing those branches that cannot contain a frequent item set [70].

II ***FP-Tree algorithm:*** overcomes the problem found in Apriori algorithm by avoiding the candidate generation process and less passes across the database, FP-Tree was found to be

faster than the Apriori algorithm. It follows a divide and conquer strategy. Firstly, it compresses the database representing frequent items into a frequent –pattern tree (FP-tree). It retains the item set association information and compressed databases are divided into a set of conditional databases, each one associated with a frequent item. It takes the help of prefix tree representation of the given database of transactions called FP tree, which saves considerable amount of memory for storing the transactions. Every node in the tree represents one item and each path represents the set of transactions that involve with the particular item. All nodes referring to the same item are linked together in a list, so that all the transactions that containing the same item can be easily found and counted [70].

We select association rule mining and classification techniques which are good for generating rules using a given dataset. The two preferred techniques are more suitable to generate rules than the other DM techniques. We combine the two forms of rules which are acquired from ARM and Classification techniques. Apriori from association rule mining and J48 decision tree algorithms from classification techniques are used to find out relevant rules that can be implemented on RBR part of the prototype system.

Generate Test Design: Prior to building a model, a procedure needs to be defined to test the model's quality and validity. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, if the test design specifies that the dataset should be separated into training and test sets, the model is built on the training set and its quality estimated on the test set.

Build Model: The purpose of building models is to use the predictions to make more informed business decisions. The most important goal when building a model is stability, which means that the model should make predictions that will hold true when it's applied to yet unseen data. Regardless of the data mining technique being used, the basic steps used for building predictive models are the same.

What happens in the build model sub phase is that we use the best data mining techniques that satisfy the business goal which is to generate rules. The first technique and specific algorithm is association rule mining and Apriori algorithm. The second technique is classification which we use specific j48 algorithm. Some sample rules generated using ARM technique are shown below

Listing 1: Rules Generated Using Apriori Algorithm

1. Appetite=Poor, Sweating=Absent and ChestIndrawing=Absent => NauseaVomitingDiarrhea=Absent
2. CoughType=Mild Cough ChestIndrawing=Absent NauseaVomitingDiarrhea=Absent => Sweating=Absent
3. Appetite=Poor, NauseaVomitingDiarrhea=Absent & CompleteBloodCount=NotChecked => ChestIndrawing=Absent
4. CoughType=Mild, Cough NauseaVomitingDiarrhea=Absent => Sweating=Absent
5. CompleteBloodCount=Not Checked Chest-X-Ray=Not Checked => ChestIndrawing=Absent

As we see in the above rules generated using ARM Apriori algorithm, the rules are not satisfying the research goal which is to make association between symptom and sign with medical diagnosis result and prescription(treatment).

The rules are only on the right side (premise) attributes of the overall dataset attributes. It means that we are not easily got the premise to direct to the conclusion part. This is because of the nature of ARM algorithm approach is unsupervised learning which is the data is drawn or divide based on how similar the data is to each other rather than drawn to specific class label attributes, But there is a probability of getting premise drawn to conclusion when we increase the number of rules generated. The following are association rules generated using Apriori algorithm when we increase the number of rules to be generated to some extent satisfying the business objective.

Listing 2: Increasing the number of Rules to be Generated Using Apriori Algorithm

1. CoughType=Dry Appetite=Poor ShortenessOfBreath=Present ==> TypesOfPneumonia=Atypical Pneumonia, Prescription=ceftroxone 19m IVBID Azithwmycin 500mg polday #3
2. CoughType=Dry Appetite=Poor ShortenessOfBreath=Present ChestIndrawing=Absent ==> Gender=Male TypesOfPneumonia=Atypical Pneumonia
3. CoughType=Dry ShortenessOfBreath=Present NauseaVomitingDiarrhea=Absent ==> TypesOfPneumonia=Atypical Pneumonia Prescription=ceftroxone 19m IVBID Azithwmycin 500mg polday #3
4. Gender=Male CoughType=Dry ShortenessOfBreath=Present ==> TypesOfPneumonia=Atypical Pneumonia Prescription=ceftroxone 19m IVBID Azithwmycin 500mg polday #3

5. Gender=Male CoughType=Dry ChestPain=Absent ==> TypesOfPneumonia=Atypical Pneumonia Prescription=ceftroxone 19m IVBID Azithwmycin 500mg polday #3
6. Gender=Female FeverLevel=High Grade Sweating=Absent NauseaVomitingDiarrhea=Absent ==> TypesOfPneumonia=Streptococcus Community Acquired Pneumonia
7. Gender=Female FeverLevel=High Grade Sweating=Absent ==> TypesOfPneumonia=Streptococcus Community Acquired Pneumonia

The second technique we use to generate rules is classification using J48 algorithm. The generated decision trees are overfitting and more complex to interpret as rules; so that we use pruning to avoid less important attributes and to minimize the complexity. The attribute Age and Prescription have many number of values so that they make the tree deeper and complex. The other attribute Chest-X-Ray is less significant; so that we remove all the three attributes. The following are some sample rules generated using J48 algorithm:

Listing 3: Rules Generated using J48 Classifier Algorithm

1. CoughType = Mild Cough, ShortenessOfBreath = Absent, ChestPain = Absent, FeverLevel = Fever: Pneumonia (49.0/2.0)
2. CoughType = Mild Cough, ShortenessOfBreath = Absent, ChestPain = Absent, FeverLevel = Low Grade: Streptococcus Community Acquired Pneumonia (50.0/2.0)
3. CoughType = Mild Cough, ShortenessOfBreath = Absent, ChestPain = Present, FeverLevel = Fever, Gender = Male: Streptococcus Pneumonia (2.0)
4. CoughType = Mild Cough, ShortenessOfBreath = Absent, ChestPain = Present, FeverLevel = Fever, Gender = Female: Streptococcus Community Acquired Pneumonia (2.0)
5. CoughType = Mild Cough, ShortenessOfBreath = Absent, ChestPain = Present, FeverLevel = Low Grade: Community Acquired Pneumonia (0.0)
6. CoughType = Mild Cough, ShortenessOfBreath = Absent, ChestPain = Present, FeverLevel = High Grade: Community Acquired Pneumonia (0.0)
7. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Absent, FeverLevel = Fever: Broncho Pneumonia (111.0/9.0)
8. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Absent, FeverLevel = Low Grade: Broncho Pneumonia (0.0)
9. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Absent, FeverLevel = High Grade: Streptococcus Community Acquired Pneumonia (1.0)
10. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Present, FeverLevel = Fever: Streptococcus Community Acquired Pneumonia (58.0)
11. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Present, FeverLevel = Low Grade: Streptococcus Community Acquired Pneumonia (0.0)

12. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Present, FeverLevel = High Grade: Streptococcus Community Acquired Pneumonia (6.0)
13. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Present, FeverLevel = NO Fever: Streptococcus Pneumonia (4.0/1.0)
14. CoughType = Dry, NauseaVomitingDiarrhea = Absent, ShortenessOfBreath = Absent, Appetite = Good: Atypical Pneumonia (7.0)
15. CoughType = Dry, NauseaVomitingDiarrhea = Absent, ShortenessOfBreath = Present: Atypical Pneumonia (121.0)
16. CoughType = Dry, NauseaVomitingDiarrhea = Nausea: Pneumonia (42.0)
17. CoughType = Dry, NauseaVomitingDiarrhea = Vomiting: Atypical Pneumonia (0.0)
18. CoughType = Dry, NauseaVomitingDiarrhea = Diarrhea: Atypical Pneumonia (0.0)
19. CoughType = Productive, FeverLevel = Fever, ChestPain = Present, Sweating = Absent: Atypical Pneumonia (2.0)
20. CoughType = Productive, FeverLevel = Fever, ChestPain = Present, Sweating = Present: Streptococcus Community Acquired Pneumonia (8.0)
21. CoughType = Productive, FeverLevel = Low Grade: Pneumonia (1.0)
22. CoughType = Productive, FeverLevel = High Grade: Streptococcus Community Acquired Pneumonia (115.0)
23. CoughType = Productive, FeverLevel = NO Fever: Community Acquired Pneumonia (40.0)
24. CoughType = NO Cough, Sweating = Absent, ShortenessOfBreath = Absent: Pneumonia (2.0/1.0)
25. CoughType = NO Cough, Sweating = Absent, ShortenessOfBreath = Present: Streptococcus Pneumonia (3.0)
26. CoughType = NO Cough, Sweating = Present: Pneumonia (9.0)

Assess Model: The model should now be assessed to ensure that it meets the data mining success criteria and passes the desired test criteria. This step is a purely technical assessment based on the outcome of the modeling tasks. Two tools commonly used to assess the performance of different models are the lift chart and the confusion matrix.

A *confusion matrix*, sometimes called a classification matrix, is used to assess the prediction accuracy of a model. It measures whether a model is confused or not; that is, whether the model is making mistakes in its predictions.

Listing 4: Summary Evaluation of Test Set

Correctly Classified Instances	322	94.152 %

Incorrectly Classified Instances	20	5.848 %
Kappa statistic	0.9285	
Mean absolute error	0.0283	
Root mean squared error	0.1342	
Relative absolute error	10.3943 %	
Root relative squared error	36.3384 %	
Coverage of cases (0.95 level)	95.9064 %	
Mean rel. region size (0.95 level)	22.6121 %	
Total Number of Instances	342	

As we see in the above listing 4, from 342 test set instances, we get 322 (94.152%) instances are correctly classified and 20 (5.848%) are incorrectly classified instances.

Table 9: Confusion Matrix for Test set

A	B	C	D	E	F	<-- classified as
49	0	1	0	0	2	a = Pneumonia
1	78	3	0	0	0	b = Streptococcus Community Acquired Pneumonia
3	3	70	0	0	1	c = Streptococcus Pneumonia
1	0	0	32	0	0	d = Community Acquired Pneumonia
0	2	0	0	43	0	e = Atypical Pneumonia
0	2	1	0	0	50	f = Broncho pneumonia

It is obvious in confusion matrix good results correspond to large numbers down the main diagonal and small, ideally zero, off-diagonal elements. As we see in table 9, confusion matrix shows large numbers down main diagonal which are indicators of good result. The small numbers (1+3+1+3+2+2+1+3+1+2+1=20) in confusion matrix are incorrectly classified. The large numbers (49+78+70+32+43+50 =322) in the diagonal of confusion matrix are correctly classified.

A *lift chart*, sometimes called a cumulative gains chart, or a banana chart, is a measure of model performance. It shows how responses, (i.e., to a direct mail solicitation, or a surgical treatment for instance) are changed by applying the model. This change ratio, which is hopefully, the increase in response rate, is called the "lift". A lift chart indicates which subset of the dataset contains the greatest possible proportion of positive responses.

4.1.5 EVALUATION

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the research objectives and seeks to determine if there is some business reason why the model is deficient. It compares results with the evaluation criteria defined at the start of the project.

In listing 3 and 4 we list out some sample rules generated using J48 classifier and Apriori association algorithms and we evaluate the rules using data mining evaluation metrics and we got good result. However the rules should be evaluated from getting the research objectives and goodness or consistent to apply in the real world application. For examples some of the rules generated are like If CoughType = NO Cough, Sweating = Absent and ShortenessOfBreath = Absent then PneumoniaType = Pneumonia (2.0/1.0). PneumoniaType = Pneumonia means that the patient are carrier of not specifically known pneumonia. But when we look for premise side of the rule the patient doesn't have any symptom of pneumonia disease. So such kinds of rules should be removed and the only logically realistic rules are selected.

A good way of defining the total outputs of a data mining project is to use the equation:

results = f(models, findings) ……Equation 1

In Equation 1, we define the total output of the data mining project as not just the models, but also the findings which can be defined as anything (apart from the model) that is important in meeting objectives of the business.

Evaluate Results: Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the data mining objectives and seeks to determine if there is some business reason why this chosen model is deficient. Another

option of evaluation is to test the model(s) on test applications in the real application if time and budget permits.

Review Process: At this point the resultant model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining project in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of Data Mining, the Review Process takes on the form of a Quality Assurance Review.

Listing 5: Filtered Rule using J48 classifier algorithm

1. CoughType = Mild Cough, ShortenessOfBreath = Absent, ChestPain = Present, FeverLevel = Fever, Gender = Male: Streptococcus Pneumonia (2.0)
2. CoughType = Mild Cough, ShortenessOfBreath = Absent, ChestPain = Present, FeverLevel = High Grade: Community Acquired Pneumonia (0.0)
3. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Absent, FeverLevel = Fever/Low Grade: Broncho Pneumonia (111.0/9.0)
4. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Absent, FeverLevel = High Grade: Streptococcus Community Acquired Pneumonia (1.0)
5. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Present, FeverLevel = Fever: Streptococcus Community Acquired Pneumonia (58.0)
6. CoughType = Mild Cough, ShortenessOfBreath = Present, Gender = Female, ChestIndrawing = Present, FeverLevel = NO Fever: Streptococcus Pneumonia (4.0/1.0)
7. CoughType = Dry, NauseaVomitingDiarrhea = Absent, ShortenessOfBreath = Present: Atypical Pneumonia (121.0)
8. CoughType = Dry, NauseaVomitingDiarrhea = Nausea: Pneumonia (42.0)
9. CoughType = Dry, NauseaVomitingDiarrhea = Vomiting: Atypical Pneumonia (0.0)
10. CoughType = Dry, NauseaVomitingDiarrhea = Diarrhea: Atypical Pneumonia (0.0)
11. CoughType = Productive, FeverLevel = Fever, ChestPain = Present, Sweating = Absent: Atypical Pneumonia (2.0)
12. CoughType = Productive, FeverLevel = Fever, ChestPain = Present, Sweating = Present: Streptococcus Community Acquired Pneumonia (8.0)
13. CoughType = Productive, FeverLevel = Low Grade: Pneumonia (1.0)
14. CoughType = Productive, FeverLevel = High Grade: Streptococcus Community Acquired Pneumonia (115.0)

Determine Next Steps: According to the assessment results and the process review, the analyst decides how to proceed at this stage. The analyst needs to decide whether to finish the project and move on to deployment or to initiate further iterations or to set up new data mining projects.

4.1.6 DEPLOYMENT

The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the client can use.

Plan Deployment: To deploy the data mining results into the business, this task takes the evaluation results and develops a strategy for deployment. If a general procedure was identified to create the relevant model, this procedure is documented here for later deployment. In this thesis the purpose of DM techniques are to produce relevant rules and to deploy the rule in the rule base part of CBRDM system.

Produce Final Report: At the end of the project, the project leader and the team write up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experiences or it may be a final and comprehensive presentation of the data mining results.

Review Project: Assess what went right and what went wrong, what was done well and what needs to be improved. When we observe some rules, they are completely out of the logic of medical diagnosis which are evaluated by domain experts. Those rules that are wrong are filtered out and only logically acceptable rules are used.

4.1 CBR PROCESS CYCLE IN CASE OF CBRDM

In the CBR cycle there are four process retrieve, reuse, revise and retain in A. amdot and E. plaza [2], but in some papers there are five process which they include representation.

4.2.1 REPRESENTATION

Knowledge can be represented using different formats. In our research we use two different knowledge representation format. Feature-vector which represent a case as feature-value pairs and if-then which used to represent rules generated using DM techniques are the two forms of knowledge representation used in our research.

Table *10:* Knowledge Represented in the Case base

	Query case	***Stored case***
Description	Age, Sex, Symptom & Sign	Age, Sex, Symptom & Sign
Solution		PneumoniaType & prescription

4.2.1 RETRIEVAL

Case retrieval is a major phase in CBR cycle where matching between two cases plays a vital role. The retrieval step is essential especially in medical applications since missing similar cases may lead to less informed decision. The reliability and accuracy of the diagnosis systems depend on the storage of cases/experiences and on the retrieval of all relevant cases and their ranking [16]. In this research we use k-nearest neighbor matching and instance-based learning algorithm. The new retrieved cases are ranked on the basis of their similarity in matching and often propose the highest ranked case as the solution of a current situation at hand [16].

4.2.2 REUSE

The third step in the case base reasoning cycle is reuse. After finding similar cases to the target problem, the system needs to reason according to the retrieved cases to find a reasonable and accurate solution for the problem. Even if there are different techniques to reuse the solution of the retrieved case [16], we prefer to apply coping the solution of the retrieved case as the solution of the target case which is easy for user of the system.

4.2.3 REVISE

After choosing to reuse a solution from the retrieved cases for a new problem, it may be discovered that this solution is, in fact, incorrect, thus providing an opportunity to learn from failure. In this phase, which is called revision, the case solution is evaluated and if the solution is incorrect, then domain specific knowledge is required to repair it [16]. In this research the domain specific knowledge is pneumonia disease diagnosis which can't easily modified or revised by the user of the system, so we use domain experts to revise the failure or incorrect cases to be stored correctly in the case base.

4.2.4 RETAIN

In this phase the new case is stored in the case base using new index and the policies of the system to add or delete case in the case base [16]. Retaining is the process of learning from the reused and revised case by users or domain experts. At the reuse phase, if the solution satisfies the user and domain experts, there is no need to go to the revise and retain phase. It is because we add the same problem and solution to the case base with only new ID number which increase the number of case in the case base which also increase the time to retrieve the case in future use of the system.

CHAPTER FIVE
IMPLEMENTATION OF THE PROTOTYPE

The framework can be downloaded from the web page: http://gaia.fdi.ucm.es/projects/jcolibri Or directly from the sourceforge.net project page: http://sourceforge.net/projects/jcolibri-cbr/. There are three main versions of the framework: v1.1, v2.0 and v2.1. we prefer to use the framework version 2.1. Actually versions 2.0 and 2.1 are very similar. jCOLIBRI 2.1 extends jCOLIBRI 2.0 including the recommenders package. Version 2.1 includes methods for developing recommendation systems. Some of the implemented methods can be used in general CBR applications and other are specific for recommender systems. Some of those General CBR methods are Filtering Retrieval method, XML utils to serialize cases and queries, Methods to obtain the query graphically (using forms), Methods to display cases, Cases retrieval using diversity and Cases selection using diversity [35].

5.1 IMPORTING JCOLIBRI2 INTO ECLIPSE

To develop a new CBR application with jCOLIBRI2; we used the Eclipse IDE. jCOLIBRI2 includes the required files to import the framework project into Eclipse: .classpath and .project. But if you import the whole project, you will have the source files of the framework into your Eclipse project. The smart solution consists on loading only the jcolibri2.jar file and related libraries into the project. That way, we will have only the source files in the Eclipse project [35].

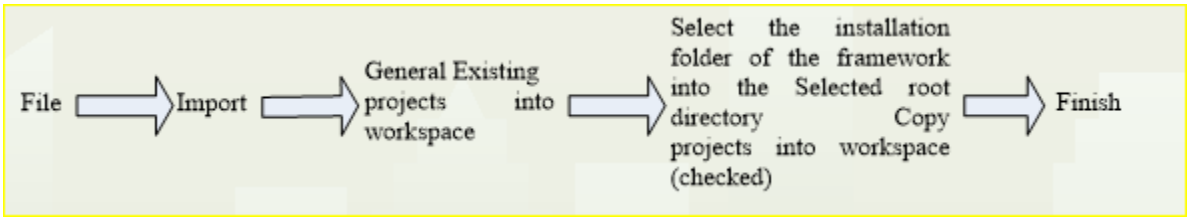

Figure 10: Import jCOLIBRI2 into Eclipse [35]

After importing jcolibri framework project, it is a good way to find out the necessary methods in order to build the CBR system.

When Constructing a Case-Based Reasoning system using the jColibri framework; we need to implement the interface jcolibri.cbrapplications.StandardCBRApplication. It requires four stages to be implemented in the main class of the system.

The four method that the interface contains are the following; configure, precycle, cycle and postcycle. In our case this class is PneumoniaDiagnosis.java, and the four stages are implemented as the main methods. The following subsections describe how we have used and implemented the methods in our system.

Listing 6: Main Class of the System

```
...
 public class PneumoniaDiagnosis extends Application implements StandardCBRApplication {

        public static CBRCaseBase _theCases;
        Connector _connector;
        public static CBRCaseBase _pneumonia;

        @FXML Button btnRetrieve; @FXML Button btnReuse; @FXML Button btnRevise;
        @FXML Button btnRetain;   @FXML AnchorPane ancPaneContent;

       public static Query queryController;
        public static Reuse reuseController;
        public static Revise reviseController;
        public static Retain retainController;
        @Override
        public void configure() throws ExecutionException {
                try {
                        // Create a database connector
                        _connector = new DataBaseConnector();
                        // Initialize the DB connector with the configuration file
                        _connector.initFromXMLfile(jcolibri.util.FileIO.findFile("databaseconfig.xml"));
                        // Create a Lineal case base for in-memory organization
                        _pneumonia = new LinealCaseBase();
                } catch (Exception e) {
                        e.printStackTrace();
```

```
                    throw new ExecutionException(e);
            }
    }
    @Override
    public CBRCaseBase preCycle() throws ExecutionException {
            // Load cases from connector into the case base
            _pneumonia.init(_connector);
            // System.out.println("connector is" + _connector);
            // Print the cases
            java.util.Collection<CBRCase> cases = _pneumonia.getCases();
            // System.out.println(cases.size() + " found");
            for (CBRCase c : cases)
                    System.out.println(c);
            return _pneumonia;
    }
    @Override
    public void cycle(CBRQuery arg0) throws ExecutionException {
                    }
    @Override
    public void postCycle() throws ExecutionException {
    }
     public void start(Stage primaryStage) throws Exception {
                                                    … }
     public void cycleButtonClicked(ActionEvent ev)
                            {
                            … }
     public static void main(String[] args) {
            PneumoniaDiagnosis s = new PneumoniaDiagnosis();
            try {
                    s.configure();
                    _theCases = s.preCycle();
            } catch (ExecutionException e) { e.printStackTrace(); }
            launch(args);
    } }
```

5.2 CONFIGURE

The configure method is the first one to be called and is responsible for preparing the system. This involves establishing a connection to the database and running the SQL script which builds the database tables and their relations. The tables are holding more than 1007 cases, complete with descriptions like the Symptom, patient Profiles, and a solution consisting of a Diagnosis Result and Recommended Prescriptions.

Furthermore, jColibri requires a database connector to be properly configured. The connector is responsible for connecting the database content to the case representation found in the system; so that the reasoning process knows which features in the case-base are which in the system. It does this through an XML document describing what classes and files are containing the description of the cases and solution.

jCOLIBRI contains three types of connectors which are JDBC Connector, TEXT Connector and DL Connector. A connector represents the first layer of jCOLIBRI on top of the physical storage. Connectors are objects that know how to access and retrieve cases from the medium and return those cases to the CBR system in a uniform way. The use of connectors give jCOLIBRI flexibility against the physical storage so the system designer can choose the most appropriate one for the system at hand.

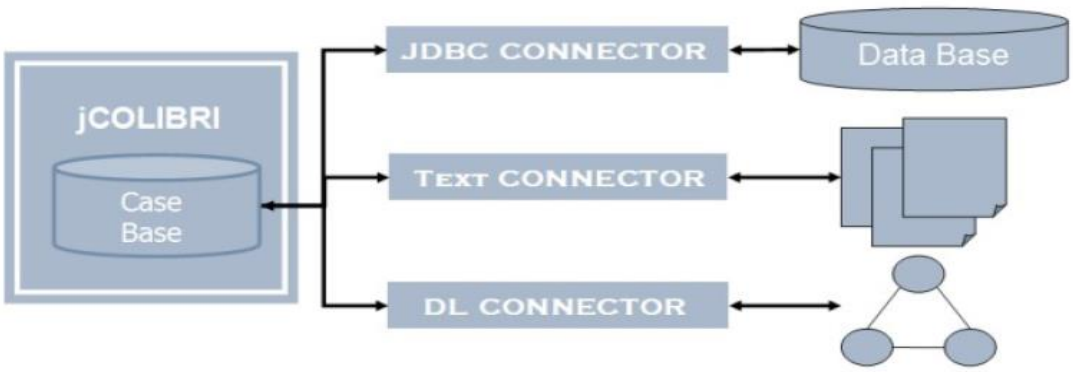

Figure 11: Case base management in jCOLIBRI2 [35]

jCOLIBRI2 includes the following connectors classes:

- jcolibri.connectors.DataBaseConnector (JDBC Connector). Manages the persistence of cases into data bases using MSSQL, MYSQL, Oracle and so on. Internally, it uses the Hibernate library.
- jcolibri.connectors.PlainTextConnector (TEXT Connector). Manages the persistence of the cases into textual files such as CSV file format.
- jcolibri.connectors.OntologyConnector (DL Connector). It uses OntoBridge2 to manage case bases stored into ontologies. OntoBridge is a sub-project of jColibri and is a library that handles the definition and use of ontologies in the CBR system. An ontology is basically a set of concepts describing the world and the relations between each such concept. The obvious interface for a connector must include methods to read the Case Base into memory and update it back into persistent media.

5.3 PRECYCLE

What happens in the precycle method is that all the cases are being loaded through the connector into the case base. Each case is reproduced with its solution attached to it, although this is just for debugging purposes. A check is done to see if every case has a solution, since there are some holes and missing data in our data set. The cases without solution are removed from the case base. The precycle iterates through the case base and loads each case into the memory, while at the same time constructing PneumoniaDescription and PneumoniaSolution objects of these, linking the concept descriptions to the classes handling each of them.

Listing 7: Java Beans Representing Problem Description

```
import jcolibri.cbrcore.Attribute;

public class PneumoniaDescription implements jcolibri.cbrcore.CaseComponent {

 String  caseId;
 String  age;
 String  gender;
 String  coughType;
 String  feverLevel;
```

```
String  appetite;
String  sweating;
String  shortenessOfBreath;
String  chestPain;
String  chestIndrawing;
String  nauseaVomitingDiarrhea;
String  completeBloodCount;
String  chest_X_Ray;
…
}
```

Listing 7 represents the problem description part of the case which is called PneumoniaDescription and in similar way in java Beans the Solution description is represented by PneumoniaSolution and contains CaseId, PneumoniaType and Prescription. Each attribute of the java beans has a getter and setter method to be accessed in java application logic.

5.4 CYCLE

When the system starts the following screen is displayed having four cycle button: Retrieve, Reuse, Revise and Retain.

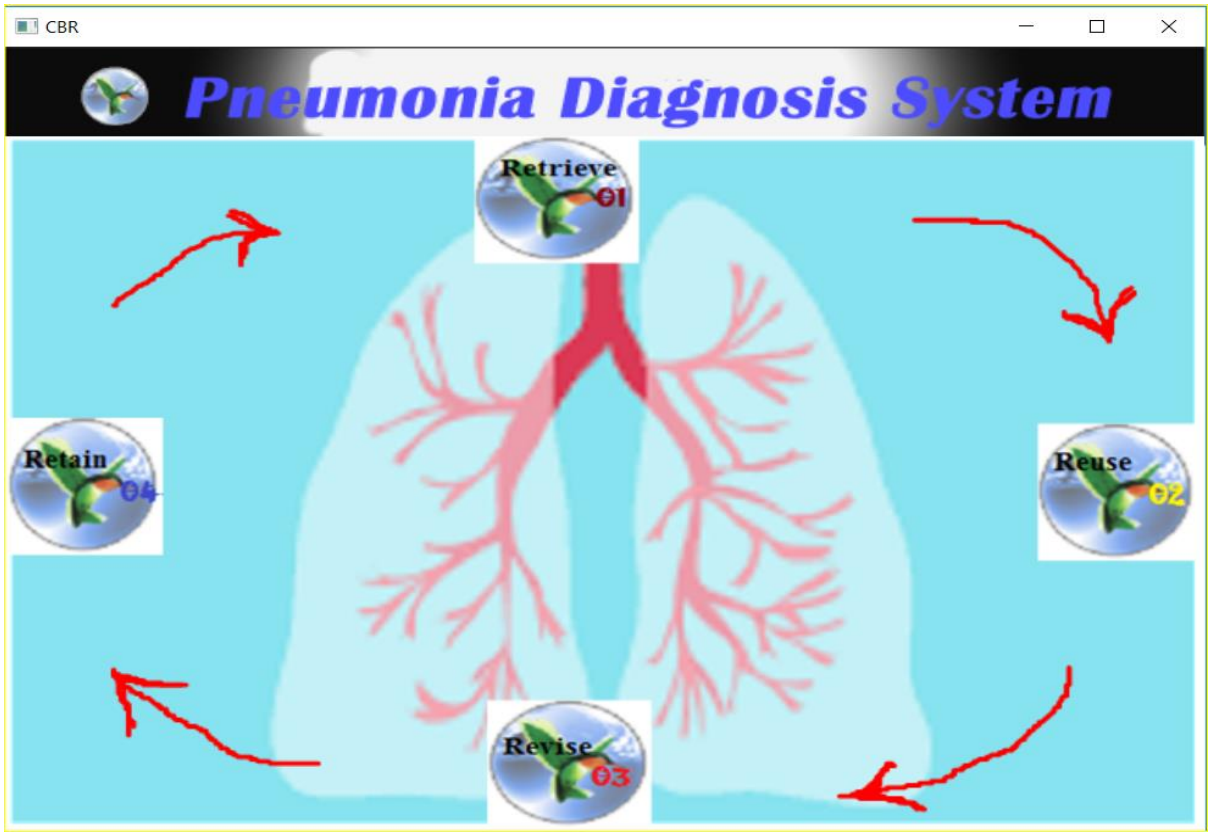

Figure 12: Screen Display when the System Start

The cycle method is the core of the system and this is where all the function of the system lies. The call is wrapped in a while-loop so that the system will be able to execute several consecutive queries without needing a reboot.

5.4.1 THE QUERY DIALOG

When we click on Retrieve cycle button the query dialog are shown. In the first query dialog we assign different values for a given attribute and there weighted values. In the sample query figure 14 below we assign age value to be 3 Year and weighted value to be 0.58, we select Female for Gender having weighted value 0.68, Dry from a list of Productive, Dry and Mild Cough for CoughType and assigning weight value 0.5. Similarly we select values for other attributes and assign weight value between 0 and 1. When we have finished selecting the right values for a corresponding attribute then we select the number of K which is the number of the most similar cases to be retrieved, finally we hit a button Send Query. The weight value is indicator of how much important the features are. For instance in our sample case, ChestPain (0.8) is a more

important feature than CoughType (0.58) which in turn is more important than Chest_X_Ray (0.23).

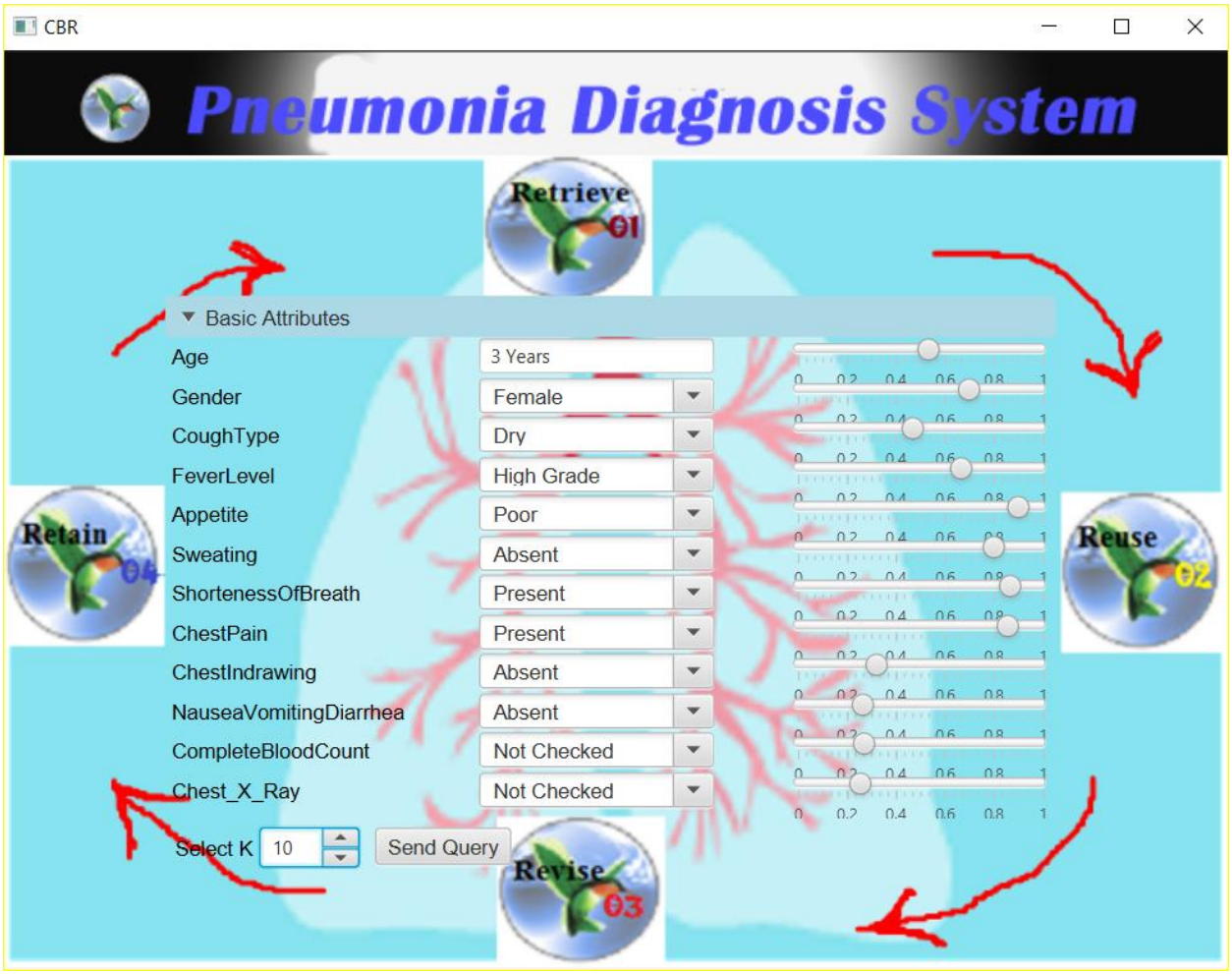

Figure 13: Query Dialog

5.4.2 THE RESULT DIALOG

In the result dialog figure 15 below, on the top there are forward and backward button to see ranked most retrieved cases. If the user has selected, e.g., 10 as the number of cases to retrieve, 10 cases are retrieved from the most relevant to the least one. There are two expandable pane parts Case Description and Case Solution. In the Case Description part the most similar problem descriptions are retrieved, in our sample case the number one most similar case description have 0.72 (72%) similarity with the original user defined case description. It has 8 attribute-value similarities out of 12 attributes.

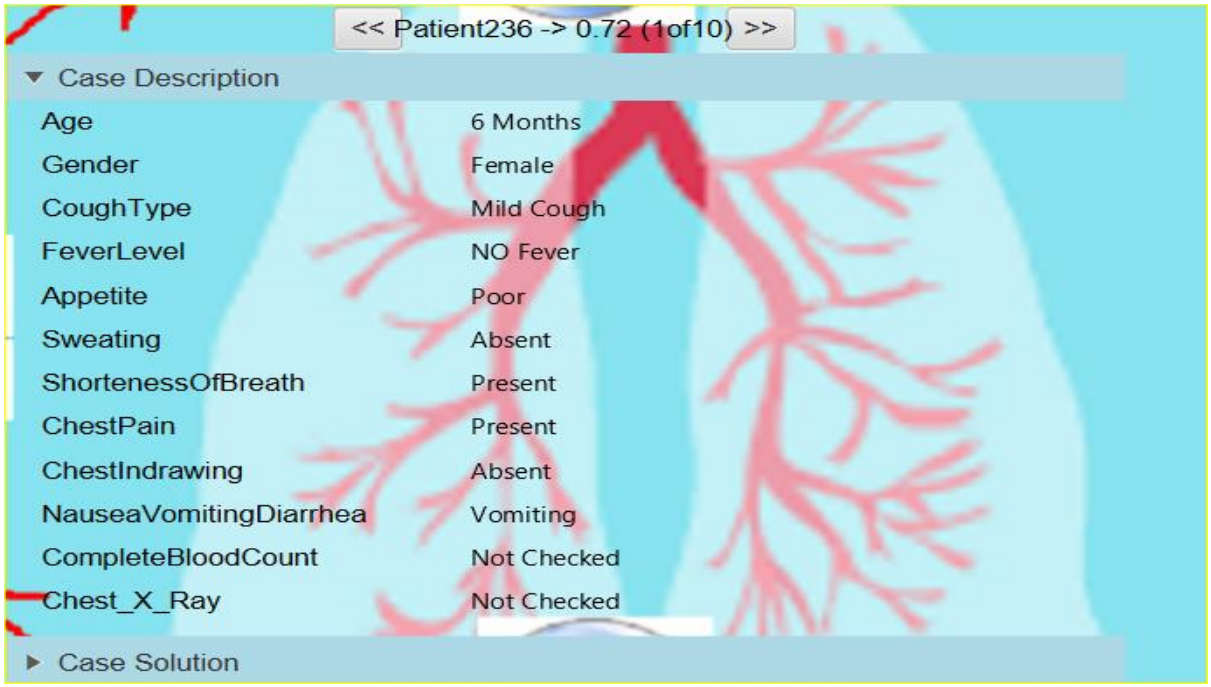

Figure 14: Result Dialog of Case Description

In the Case Solution part the corresponding solution will be retrieved and displayed. It has two attributes Pneumonia Type and Prescription.

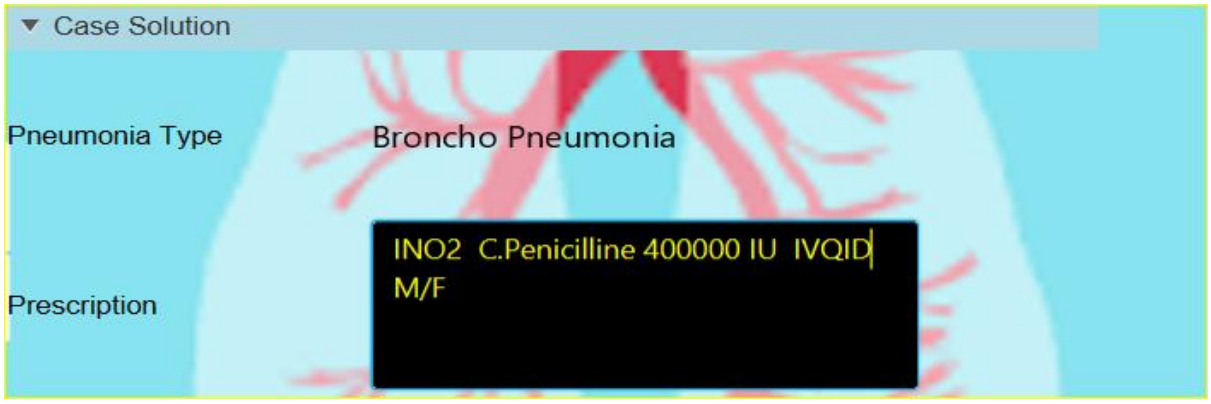

Figure 15: Result Dialog of Case Solution

The CBR part of the system evaluates each case accordingly and uses a pretty standard k-Nearest Neighbor function to retrieve the top k cases from the case-base. The two methods or functions are NNScoringMethod.evaluateSimilarity SelectCases.selectTopK.NNScoringMethod.evaluateSimilarity computes the similarity between

the query and each case in the case base. These are stored in a Collection of cases and passed to the solution dialog and then presented to the user.

5.4.3 THE REUSE DIALOG

When we click on the reuse cycle button only the solution part is displayed to use by the user of the system. There are two techniques for reuse process which are Direct Copying method and Adaptation method. In a Direct Copying method the differences are abstracted away (they are considered non relevant while similarities are relevant) and the solution class of the retrieved case is transferred to the new case as its solution class.

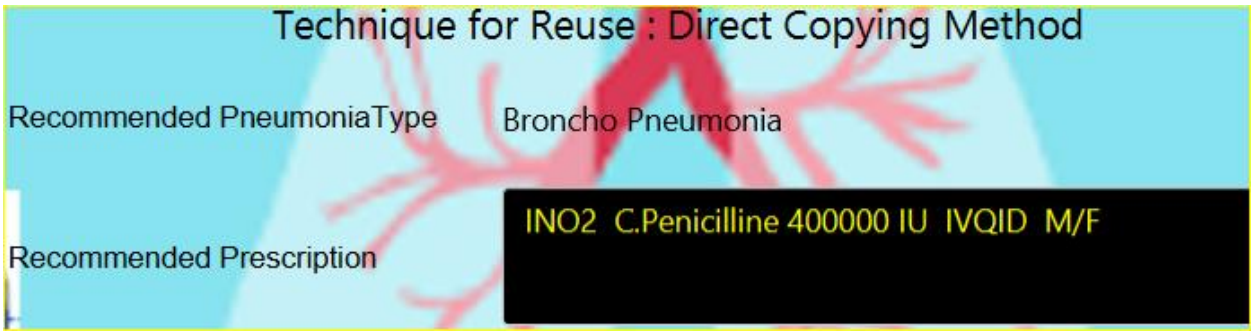

Figure 16: Reuse Dialog

5.4.4 THE REVISION DIALOG

When a case solution generated by the reuse phase is not correct, an opportunity for learning from failure arises. This phase is called case revision and consists of two tasks: (1) evaluate the case solution generated by reuse. In our case the solution is evaluated by domain experts using his/her pneumonia diagnosis experiences. If successful, learning from the success (case retainment), (2) otherwise repair the case solution using domain-specific knowledge. For example in our specific case sample the user is 3 Years old, but the solution retrieved to be used in the reuse phase is for 6 month infant; so that in the Prescription part of the solution the dosage of the medicinal drug changed for 3 years old girl and in the case description age modified to be 3 Years girl.

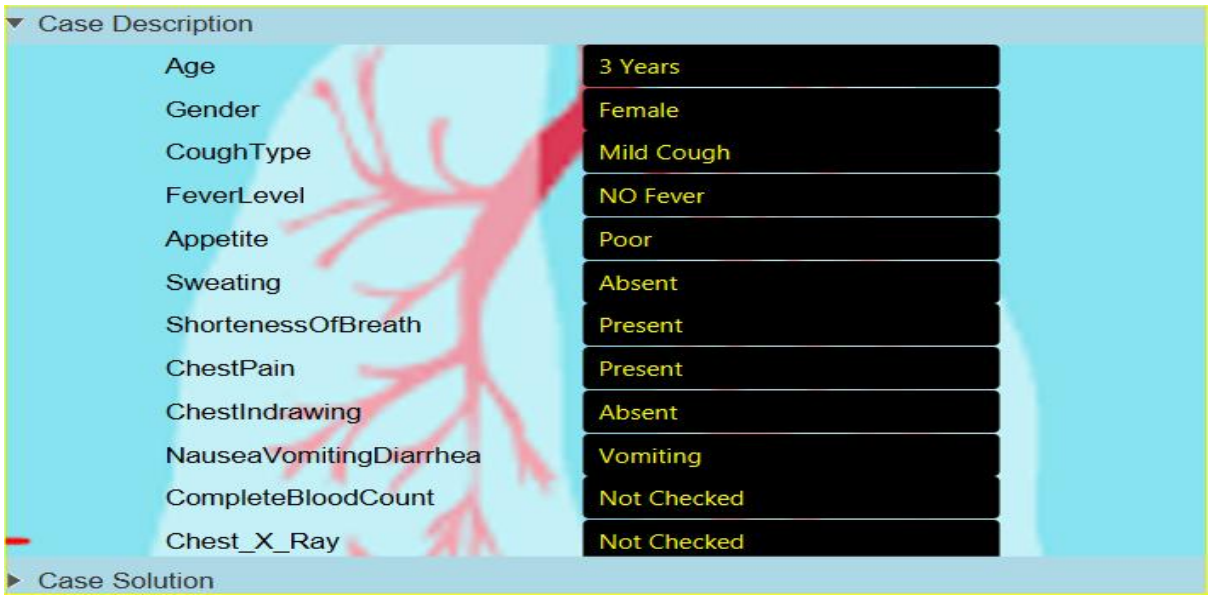

Figure 17: Revision Dialog of Case Description

In the solution part the prescription retrieved for 6 month infant is INO2, C. Penicillin (Crystalline penicillin) 400000 IU IVQID, M/F and changed to INO2, C. Penicillin(Crystalline penicillin) 800000 IU IVQID, M/F for 3 Years old girl. This is modified by domain expert and as we see in the figure below the only change in the case solution is the dosage to be 800000 IU.

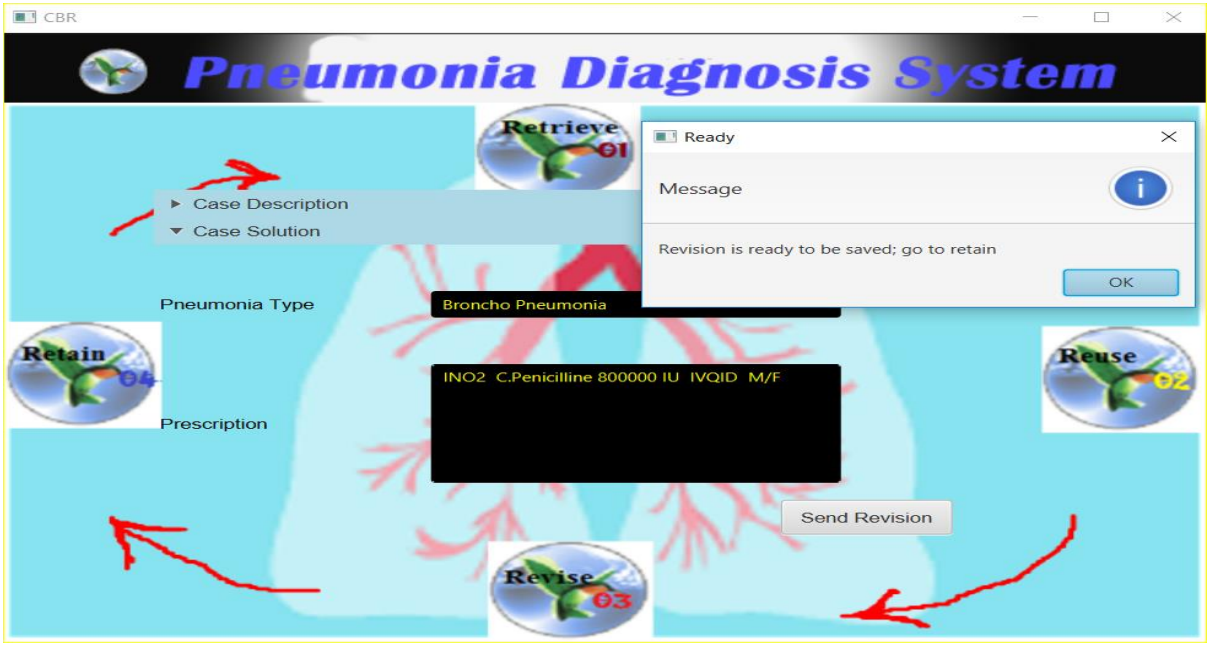

Figure 18: Revision Dialog of Case Solution

5.4.5 THE RETAIN DIALOG

In this step, the new case will be added to the case base according to some policies in the system. Retention includes adding knowledge and new cases to the case base, all which needs to be indexed, as well as deleting cases from the case base in order to restrict its growth. When we click save button the new case will be added to the case base with new CaseId. For example in our specific case what is modified by domain expert is dosage to be 800000 IU and the age of the girl to be 3 Years. Finally when we click save button, it will be save as modified by expert.

5.5 POST-CYCLE

It disconnects the connector to the database and effectively shuts down the system by closing the connector and database server. If additional steps are needed, all clean-up may be performed in this method.

5.6 INTEGRATION OF RULES GENERATED USING DM AND CBR

The concept behind the integration of data mining and case based reasoning is to use data mining for the purpose of preprocessing and generating the rules where that rules is used to build rule base reasoning and the preprocessed dataset is used for developing the CBR prototype system. The need for such kinds of integrations are to reach to make strong reasoning towards user defined query (case description). The reasoning mechanism of CBR is supported by the general rule generated by the DM Techniques.

Table 11: Integration Of DM and CBR

CBR		***Rules form DM***	
Case Description	***Case Solution***	***Premise (If)***	***Conclusion (Then)***
Age= 30 Year Gender=Male, CoughType=Dry, FeverLevel=Fever, Appetite=Poor, Sweating=Absent, ShortenessOfBreath = Present, [CaseId=Patient31, 0.74]	TypesOfPneumonia =Atypical Pneumonia Prescription=ceftroxone 19m IVBID Azithwmycin 500mg polday #3	Gender=Male **And** CoughType=Dry **And** ShortenessOfBreath= Present [Rule #4 from listing 2]	TypesOfPneumonia=Atypical Pneumonia Prescription=ceftriaxone 19m IVBID Azithwmycin 500mg polday #3
Age= 25 Year Appetite=Good CoughType=Dry ShortenessOfBreath =Present NauseaVomitingDiarrhea=Nausea [CaseId=Patient237, 0.75]	TypesOfPneumonia =Pneumonia Prescription=ceftriaxone 19m IVBID Azithwmycin 500mg polday #3	CoughType=Dry **And** ShortenessOfBreath= Present **And** NauseaVomitingDiarrhea=Absent [Rule #3 from listing 2]	TypesOfPneumonia=Atypical Pneumonia Prescription=ceftriaxone 19m IVBID Azithwmycin 500mg polday #3
CoughType =Productive, ChestPain = Present, ShortenessOfBreath =Absent, FeverLevel= High Grade [CaseId=Patient156, 0.74]	TypesOfPneumonia =Streptococcus Community Acquired, Prescription= INO2 Azithromycin 500mg polday Ceftvaxone 19m IVBID	CoughType = Mild Cough **And** ShortenessOfBreath = Absent **And** ChestPain = Present **And** FeverLevel = High Grade [Rule #2 from Listing 5]	TypesOfPneumonia =Community Acquired Pneumonia

Age= 10 Month, Gender = Female, CoughType = Mild Cough, ShortenessOfBreath = Present, ChestIndrawing = Absent, FeverLevel = NO Fever [CaseId=Patient8, 0.68]	Crystalin penicillin 750 000 IU IVQID INO2 Paracetamol suppose 125mg PRN , TypesOfPneumonia =Streptococcus Pneumonia	CoughType = Mild Cough **And** ShortenessOfBreath = Present **And** Gender = Female **And** ChestIndrawing = Present **And** FeverLevel = ¬ Fever [Rule #6 from Listing 5]	TypesOfPneumonia =Streptococcus Pneumonia

As we see in the table 11 above, the number of attributes used by DM part of Premise (If) are three or four to reach to conclusions (Then), but the CBR use all the 12 attributes to recommend the solution. However, the two techniques used different number of attributes they reach to more or less similar solution or conclusion. The general rule produced by DM techniques support the specific case of CBR paradigm. Even if the RBR part of the prototype are not implemented, it works in supporting each other just like in table 11 above.

CHAPTER SIX

EVALUATION, RESULT AND DISCUSSION

This chapter contains the evaluation done on our CBRDM system. A good way of evaluating a CBR system is to get an expert within the domain and examine the retrieved cases to evaluate if the results are satisfying or not. We have performed this by checking if the system returns the right cases by taking query cases which already exists in the database.

6.1 SYSTEM PERFORMANCE TEST

The system performance test is done using two important metrics: The first metrics is Accuracy (similarity of the user Query with Retrieved case) and Retrieval performance (using Recall and Precision to retrieve case solution from the case base)

6.1.1 ACCURACY

The result variable is the similarity between the best matching case and query, 1 to 0.This number is added to a data series which is used to create a graphical representation of the overall similarity for all cases.

$$\sum_{i=1}^{n} w_i * sim(f_i^l, f_i^R) \; / \; n \sum_{i=1}^{n} w_i \quad \text{.....Equation 2}$$

Where:

w_i : is the importance of the feature (slot) i

Sim() : is the similarity function

f_i^l, f_i^R : are the values for feature $f\ i$ in the source and target cases, respectively

n : is the number of attributes in each case

The similarity function is defined as follows:

$$sim(f_i^l, f_i^R) = 1 - (| f_i^l, f_i^R | \; / \; | f_{max} - f_{min} |) \quad \text{.....Equation 3}$$

The jcolibri uses the KNN algorithm to retrieve the top k similar cases with the users problem, the above equation 2 and 3 are used by jcolibri framework of KNN algorithm to retrieve the top K most similar case solution. To evaluate the accuracy we randomly formulate 20 queries and test them on the prototype system and the result presented as shown in table 11.

Table 12: Accuracy Performance Evaluation

NO	*Case ID*	*Similarity Values*	*NO*	*Case ID*	*Similarity Values*
1	Patient137	0.78	11	Patient95	0.60
2	Patient238	0.76	12	Patient156	0.67
3	Patient78	0.57	13	Patient994	0.62
4	Patient238	0.74	14	Patient38	0.54
5	Patient100	0.66	15	Patient5	0.64
6	Patient8	0.61	16	Patient8	0.73
7	Patient16	0.59	17	Patient78	0.64
8	Patient15	0.76	18	Patient1001	0.53
9	Patient133	0.68	19	Patient38	0.54
10	Patient232	1.0	20	Patient876	0.60
Average Similarity			**0.67**		

As we see in the above table 12 the average similarity value is **67%** which is above good value and our domain area is pneumonia disease diagnosis; so what is expected is to be high accuracy to confidently recommend to use the system.

6.1.2 EVALUATION OF THE RETRIEVAL PROCESS

The standard metrics to measure the performance of retrieval process in CBR system is to use precision and recall. According to [71], Recall is the ratio of the number of relevant cases retrieved to the total number of relevant cases in the case base and precision is the ratio of the number of relevant cases retrieved to the total number of irrelevant and relevant cases retrieved. These are usually expressed as a percentage.

To perform the evaluation, for each test case the relevant pneumonia cases from the case base should be identified. For identification of relevant cases, test cases are given to the domain expert in order to assign possible relevant cases from the testbed to each of the test cases. The domain expert uses the value of pneumonia type and Recommended prescription (Solution) attributes of the pneumonia case as the main concept to assign the relevant case to the test cases. After the identification of the relevant cases to the test cases by the domain expert, precision and recall are calculated.

Table 13: Relevant Cases Assigned by Domain Experts for the Sample Test Cases

Test Case	***Relevant cases from the case base***	***#Case [1.0,0.8]***
Patient990	Patient1003, Patient945, Patient954, Patient958, Patient994, Patient156	5
Patient10	Patient232, Patient237, Patient251, Patient266, Patient289, Patient308, Patient366, Patient401, Patient414, Patient424,	10
Patient111	Patient85, Patient119, Patient122, Patient143, Patient153, Patient102, Patient22, Patient144	6
Patient32	Patient539, Patient545	1
Patient399	Patient539, Patient150, Patient159, Patient167, Patient215, Patient227, Patient267, Patient286	7
Patient209	Patient271, Patient297, Patient338, Patient357, Patient388, Patient430	5

According [72], there is no standard threshold for degree of similarity that has been used for retrieving relevant cases. That is why different researchers use different case similarity threshold to measure the performance of their system. We use the commonly used threshold interval [1.0,0.8].

The next step is to find out retrieved relevant cases from the case base by using the threshold interval [1.0, 0.8] which is commonly used interval. We use the test case query to retrieve the 10

best cases from the case base. The number 10 is K which the CBR system retrieve top 10 case but its relevance is provided if it is in the give threshold interval. After we get the number of relevant retrieved cases in the threshold interval, we can calculate precision and recall.

Table 14: Recall and Precision results for the sample test cases

Test Case	***Recall***	***Precision***
Patient990	0.83	0.5
Patient10	1.0	1.0
Patient111	0.75	0.6
Patient32	1.0	0.2
Patient399	0.87	0.7
Patient209	0.84	0.5
Average	**0.88**	**0.58**

The table 14 above, shows the calculated value of recall and precision for each sample case and their average value. The average recall and precision value of the system is **88%** and **58%** respectively. The higher recall value shows that the system obtains most of relevant cases from the case base. Therefore, this CBR system can retrieve relevant cases that enable users to make decision easily in pneumonia disease diagnosis. On the other hand, the system retrieved relevant cases to the system with 58% precision. The precision value of the system is not as expected by the researcher due to few number of cases used. As the number of cases increased, the precision value of the system will also increase and better performance will scored in retrieving relevant cases. When we look for recall, it is more than very good result and precision is approximately good result.

6.2 ACCEPTANCE TEST

User acceptance test objectives are to demonstrate that the prototype software system satisfies customer requirements. The development of the acceptance criteria should occur when the requirements for the software system are initially defined. The acceptance criteria should specify baselines against which the performance of the functional requirements and other features and deliverables of the software system will be evaluated.

Parameters of evaluation criteria for user acceptance test are adapted from [73] and checked to fulfill the standard metrics of software quality from user perspectives. We prepared eight Parameters of evaluation criteria questions for user of prototype system with five performance value for analyzing the prototype system: Poor=1, Fair=2, Good=3, Very good=4 and Excellent=5. The eight evaluation criteria questions are checked to fulfill the IEEE standard metrics of User acceptance test. The IEEE standard metrics are listed below in [33] satisfies the research software acceptance from user perspective interview questions in table [15].

Functionality, Maintainability, Testability, Ease of installation, Standards, adherence, Completenes, Performance, Adequacy of documents, Flexibility, Integrity, Reliability ,Usability, Portability, Security

6.2.1 PROTOTYPE SYSTEM ACCEPTANCE EVALUATION BY DOMAIN EXPERT

We select 6 domain experts in the pneumonia disease diagnosis to make CBRDM system acceptance test by domain experts.

Table 15: Acceptance Evaluation by Domain Experts

NO	***Parameters of Evaluation Criteria***	***Performance Value***					
		1	**2**	**3**	**4**	**5**	***Average***
1	Easy to use of the case based reasoning system			1	3	2	**4.16**
2	Is the system efficient in time			2	4		**3.66**
3	Is the user interface interactive				4	2	**4.33**
4	Adequacy and clarity of decision support			1	2	3	**4.33**
5	Relevancy of the retrieved case in the decision making				3	3	**4.5**
6	Fitness of the final solution to the new case			1	5		**3.83**

7	Relevancy of the attributes in representing patient case			1	3	2	**4.16**
8	Rate the significance of the system in the domain area				2	4	**4.66**
	Average						**4.20**

Average Performance (Given Criteria) = $\sum_{i=1}^{n=5} PVi * {}^{\#R}/_{Total\#R}$……..…Equation 4

Where

PVi --------- Performance Value

#R --------- Number of Respondents

Total #R --------- Total Number of Respondents

We can calculate average performance of a given criteria for domain expert acceptance test by using the above equation 4. For instance in the above table 15, Average Performance for Easy to use of the CBR system (1) criteria is ((**1***0+**2***0+**3***1+**4***3+**5***2)/6= **4.16**), the bold numbers are performance value given by the correspond number of respondents and 6 is total number of respondents in user acceptance. In the same way we can calculate average performance for other criteria.

The table 16 below show percentage of respondents rate for a given parameter of evaluation criteria. We can interpret the table 16 below by looking for criteria in the above table 15 and its corresponding order number matching with the table 16 order number below.

For instance for order **NO 1**, the performance evaluation criteria is easy to use of the CBR system and we can interpret it as follows, **16.7%** of the respondents rate easy to use of the case based reasoning system as good, **50%** of respondent rate as very good and the remaining **33.3%** of the respondent's rate it as excellent. All the other criteria are interpret in the same way.

Table 16: Percentage of Respondents Rate By Domain Experts

NO	*POOR (%)*	*FAIR (%)*	*GOOD (%)*	*VERY GOOD (%)*	*EXCELLENT (%)*	*Average (%)*
1	0.00	0.00	16.7	50	33.3	**83.2**
2	0.00	0.00	33.3	66.7	0.00	**73.2**
3	0.00	0.00	0.00	66.7	33.3	**86.6**
4	0.00	0.00	16.7	33.3	50	**86.6**
5	0.00	0.00	0.00	50	50	**90.0**
6	0.00	0.00	16.7	83.3	0.00	**76.6**
7	0.00	0.00	16.7	50	33.3	**83.2**
8	0.00	0.00	0.00	33.3	66.7	**93.2**
					Total Average (%)	**84.0**

In table 16 above, the average percentage shows percentage of prototype system acceptance by domain experts for specified parameter of evaluation criteria. For example for **NO 8 (**Rate the significance of the system in the domain area which are the criteria in table 15) is accepted by domain experts **93.2 %** which is above Very good and approximately to excellent. Similarly, we can interpret all the other criteria for average prototype system acceptance by domain experts. In general for all criteria the developed prototype system is accepted by domain experts **84 %** which is above very good. This performance result shows the prototype has a promising applicability in pneumonia disease diagnosis.

6.2.2 COMPARISON OF THE PERFORMANCE OF CBRDM WITH PREVIOUS CBR SYSTEMS

Average similarity of cases and acceptance test comparison are done in this research. Similarly, the recall and precision value of retrieval performance of the systems developed by the earlier thesis research are compared as shown in table 17 below.

Table 17: Comparison of the Performance of CBRDM with Previous CBR Systems

Domain and Researcher	***Tools***	***Retrieval Performance***		***Average Similarity of cases***	***Acceptance Test***
		Recall	***Precision***		
Mental Health (Getachew Wassie, 2012)	jCOLIBRI 1.1	82%	71%	74.5%	83.2%
Investment Sector (Yibeltal Chanie, 2013)	jCOLIBRI 1.1	85%	64%	87%	82%
Field of study selection Biazen(2013)	jCOLIBRI 1.1	85%	55%	Not evaluated	77.2%
Pneumonia Diagnosis	jCOLIBRI 2.1	88%	58%	67%	84%

As indicated in table 17, the result of the recall value of the system shows a little bit enhancement from Getachew [73], Biazen [74] and Yibeltal [75], while the value of precision shows decline from the precision value of Getachew [73] and Yibeltal [75], but it shows a little improvement from Biazen [74]. On the other hand, Average Similarity of cases of the system go down from Getachew [73] and Yibeltal [75], but Biazen [74] does not evaluate it.

The other comparisons done is acceptance evaluation of CBRDM prototype with other CBR systems. All those acceptance evaluation are done from domain expert perspectives and as indicated in the table 17, we have betterment from the entire three-mentioned researcher's domain expert acceptance evaluation. The framework used in our research is the latest version of jcolibri 2.1, but all the other researchers used jcolibri 1.1 version.

CHAPTER SEVEN
CONCLUSIONS AND RECOMMENDATIONS

Nowadays, the application of AI in medical domain interests many researchers especially hybriding the sub field of CBR with other AI or ML Techniques for medical diagnosis. The main focus of this research is diagnosing pneumonia using CBR and DM approach. Therefore, this chapter concludes the overall findings of this research and based on the findings recommends the remaining problems to be investigated by other researchers.

7.1 CONCLUSIONS

The general objective of this research is to enhance the effectiveness and efficiency of pneumonia disease diagnosis mainly from increasing the accuracy and retrieval performance. To do this, what has been demonstrated in this report is basically how and why the combination of CBR and DM can work well in a diagnosis system. In health science it is important to have a strong foundation for every action taken toward treatment of patients, as it is their health and possibly their lives which are at stake. Many successful hybrid decision support systems and recommendation systems are already in wide-spread use today, but the integration of CBR and DM Techniques are rarely researched and implemented. It is vital to utilize the full strength of a combination of two paradigms, and therefore we believe that our system has something to offer.

The strength of the system comes from the fact that both previous experiences and dataset from data mining analysis work together to provide the best possible solution. Where, CBR handles certainty and facts collected from expert sources, the DM deals with producing a general knowledge to strengthen those specific cases. The first is a good representation of the world as it have been experienced already, while the latter provides rules as an association of symptoms and sign to the diagnosis result and doctors prescriptions.

Even though the idea behind the design is good, the implementation presented still needs further work and experimentation in order to be viable for use in larger and more complex domains. The solution presented shows one option in how to give the user full access to and insight into the

system and its processes. The method of modeling the world in a causal network is not a new one, neither is providing solutions in the form of old cases. The combination of the two however, might be a new one, and as the results show, the system is capable of achieving good results even with much uncertainty in its model.

Based on the result obtained from respondent's evaluation points, the proposed prototype for CBRDM the total average performance of the domain expert acceptance evaluation is about 84.00% of the proposed system prototype satisfied by domain experts. Similarly, the total system performance testing using test cases were obtained about 67% of proposed system prototype was accuracy tested. The retrieval performance of the system are evaluated using recall and precision. The result of recall is about 88% and its precision is 58%. This results shows how the domain experts satisfied with proposed CBRDM and they hope the developed prototype helps to solve domain problems.

Along with the human-reasoning based method of CBR, the combination of these two paradigms will with more work lead to powerful explanation facilities in decision support systems. However, it would be interesting to try to combine the DM explanations with more complex CBR based explanations as this is not something done in this report. The possibility for a more knowledge-intensive solution like a semantic net representation of the cases together with the causal properties of the DM should possibly lead to some interesting solutions, which in the end would only benefit the users of these decision support systems.

The primary dataset source is Adama hospital and medical college and also we acquire the data from different standard guidelines and general practitioners manuals. All the data is collected using a document review techniques rather than interviewing a domain experts. The dataset has noise data that needs to be prepared and preprocessed to avoid and clear-cut those noise data. To do this a data mining tool weka are used for statistical calculations. The preprocessed data set are used as a case in the case based reasoning and further used to generate rules using classification and association rule mining techniques. We use DM techniques to produce rules instead of human experts which are cost and time consuming processes. Even if, we are not apply the attribute selection techniques of data mining, it is very important to find out relevant attributes from a set of attributes by looking for high frequency of an itemset. Instead of human expert to give weight

for relevant attributes, it is worthy to use data mining techniques to put attribute rank by looking for its high frequency. To do this a lot of records are necessary in the dataset.

The significance of the system from the domain expert view is unquestionable. From domain expert acceptance test result, we got an average about 93% acceptance of the parameter rate the significance of the system on the domain area which indicate such kinds of clinical decision support system are very important. The CBRDM system can minimize the workload of professional health workers, general practitioners can get a lot of lessons/training from the system and also the patients can be supported by the system.

7.2 RECOMMENDATIONS

- The system focus on a single pneumonia disease diagnosis system, but there are many respiratory tracts infections disease that probably have similar symptoms. So we recommend for the system developer to develop diagnosis system that include all respiratory tracts disease.
- One challenge in disease diagnosis/recommender system is data collection process. As we know clinical data are so secured and are not allowed to be given to anybody. The data is stored in medical record cards rather than in softcopy/digital format in computer. So it is bulky to collect all the necessary data from medical record cards. So we recommend to data owner to store their medical data in digital database and to use electronic medical record system.
- In our country there is lack of human experts in many fields, especially in health area. So to support those professionals, to give service (recommendation) to the users and to train practitioners in the health field. We need more expert systems and many research to improve the efficiency and effectiveness of these expert system.
- To develop robustness system that probably contain all patient related problems that include suggesting diagnosis result, giving prescribed medicine, giving price for prescribed medicine etc.
- In our thesis the architecture is to integrate DM, CBR and RBR, but in the implementation of the prototype shows only DM and CBR. So we recommend any researcher to make the full prototype for the betterment of such kinds of hybrid system.

REFERENCES

[1] Kim Ohme Pedersen, explanation methods in clinical decision support: a hybrid system approach, norwegian university of science and technology, norwegian, 2010.

[2] A. Aamodt, E. Plaza, Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications. IOS Press, Vol. 7: 1, pp. 39-59, 1994.

[3] Jay Liebowitz, The Handbook of Applied Expert Systems, CRC Press LLC.

[4] David L. Olson, Dursun Delen, Advanced Data Mining Techniques, 2008 Springer-Verlag Berlin Heidelberg.

[5] http://www.healthline.com/health/pneumonia#Overview1 n.d.

[6] Pneumonia and diarrhoea: Tackling the deadliest diseases for the world's poorest children. New York, UNICEF, 2012.

[7] Levels and Tends in Child Mortality: Report 2011. UN Inter-agency Group for Child Mortality Estimation, 2011.

[8] Rudan I. et al. Setting research priorities to reduce global mortality from childhood pneumonia by 2015. PLoS Med, 2011, 8(9):e1001099.

[9] http://www.nationmaster.com/country-info/stats/Health/Physicians/, 2003-2018.

[10] Jim Prentzas and Ioannis Hatzilygeroudis, Categorizing approaches combining rule-based and case-based reasoning, Department of Computer Engineering and Informatics, School of Engineering, University of Patras, 26500 Patras, Greece.

[11] https://www.unicef.org/esaro/5440_eth2014_pneumonia.html/ Retrieved date 09/02/2018

[12] Salama A. Mostafa, et al, A Soft Computing Modeling to Case-based Reasoning Implementation, International Journal of Computer Applications (0975 – 888) Volume 47– No.7, June 2012

[13] Yiyu Yao, Ning Zhong et.al, A Conceptual Framework of Data Mining, Data Mining: Foundations and Practice, SCI 118, 501-515, Springer-Verlag, 2008.

[14] Azeb Bekele, Integrated Case Based and Rule Based Reasoning for Decision Support, Norwegian University of Science and Technology Department of Computer and Information Science, July 2009.

[15] Ralph Bergmann et.al, Representation in case-based reasoning, The Knowledge Engineering Review, Vol. 00:0, 1–4.c 2005, Cambridge University Press.

[16] Sima Soltani, Case-Based Reasoning for Diagnosis and Solution Planning, Technical Report No. 2013-611, Queen's university, Kingston, Ontario, Canada, October 2013.

[17] Vinay Tiwari, International Journal of Computer Science & Information Technology (IJCSIT), Vol 2, No 6, December 2010.

[18] Ernest Friedman-hill (2003), Jess in Action: Rule-Based Systems in Java, United States of America, Manning Publications Co.

[19] Shweta Tyagi and Kamal K. Bharadwaj, A Hybrid Recommender System Using Rule-Based and Case-Based Reasoning, International Journal of Information and Electronics Engineering, Vol. 2, No. 4, July 2012.

[20] Mariana Cabrera and Ernesto Edye, Integration of Rule Based Expert Systems and Case Based Reasoning in an Acute Bacterial Meningitis Clinical Decision Support System, International Journal of Computer Science and Information Security, Vol. 7, No. 2, 2010.

[21] Sima Soltani, Case-Based Reasoning for Diagnosis and Solution Planning, Technical Report No. 2013-611, Queen's university, Kingston, Ontario, Canada, October 2013.

[22] Isabelle B. and Amalia A. , Case Based Reasoning with Bayesian Model Averaging: An Improved Method for Survival Analysis on Microarray Data, University of Washington Tacoma, Institute of Technology USA, 2010.

[23] Shaker El-Sappagh, Mohammed Elmogy, et al, A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis, Artificial Intelligence in Medicine 65 (2015).

[24] Sangyong Kim and Jae Heon Shim, Combining case-based reasoning with genetic algorithm optimization for preliminary cost estimation in construction industry (www.nrcresearchpress. com/cjce on 5 November 2013.

[25] Dr. Trupti P. Shah and Pooja J. Shah,(2013), Connectionist Expert System for Medical Diagnosis using ANN- A case study of skin disease Scabies, IJARCSSE, Univesity of Baroda, Vadodara, India, Volume 3, Issue 8, August 2013 ISSN: 2277 128X.

[26] Stefano Bragaglia, Federico Chesani, et al, An Hybrid Architecture Integrating Forward Rules with Fuzzy Ontological Reasoning, June 2010, University of Bologna, Bologna, Italy.

[27] Jia Tian, et al, Multi-Modal Reasoning Medical Diagnosis System Integrated With Probabilistic Reasoning, International Journal of Automation and Computing 2 (2005).

[28] Joyce Jackson, Data mining: A Conceptual Overview, Communications of the Association for Information Systems (Volume 8,2002).

[29] Isabelle Bichindaritz, Data Mining Methods for Case-Based Reasoning in Health Sciences, In Proceedings of the ICCBR 2015 Workshops. Frankfurt, Germany.

[30] Nikola K. Kasabov(1998), Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering , The MIT Press Cambridge, Massachusetts London, England.

[31] http://www.generation5.org/content/2005/PDAMum.asp

[32] Ian H. Witten, et al (2011), Data Mining Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, USA.

[33] N. Juristo and J.L. Morant, "Common Framework for the Evaluation Process of KBS and Conventional Software," Knowledge Based Systems," vol. 11, pp. 145-159, October 1998.

[34] Armin S. and Thomas R, Rapid Prototyping of CBR Applications with the Open Source Tool myCBR, German Research Center for Artificial Intelligence (DFKI) GmbHm.

[35] Juan A. Recio-García et.al, jCOLIBRI2 Tutorial, Group for Artificial Intelligence Applications, September 16, 2008, De Madrid, Spain.

[36] Nasrullah I. and Muhammad H Ashraf, Evaluation of jCOLIBRI, Malardalen University, Sweden, 2006.

[37] A. Atanassov, L. Antonov, Ccomparative Analysis of Case based Reasoning Software Frameworks jCOLIBRI AND myCBR, Journal of the University of Chemical Technology and Metallurgy, 2012.

[38] Blaz Zupan, et al (2008), Open-Source Tools for Data Mining, Baylor College of Medicine, Houston, TX, USA.

[39] Morteza Behbahani, et al, A case-based reasoning system development for statistical process control: Case representation and retrieval, Computers & Industrial Engineering 63 (2012) 1107–1117

[40] Abdel-Badeeh M. Salem, Case Based Reasoning Technology for Medical Diagnosis, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering Vol:1, No:7, 2007.

[41] Alizadehsani, et al., A data mining approach for diagnosis of coronary artery disease, Comput. Methods Programs Biomed, 2013.

[42] Abdel-Badeeh M, et al, a case based expert system for supporting diagnosis of heart disease, AIML Journal, Volume (5), Issue (1), March, 2005 R.

[43] Negny Ste´phane, et al, Effective retrieval and new indexing method for case based reasoning: Application in chemical process design Engineering, Applications of Artificial Intelligence 23, 2010.

[44] Shivam Agarwal, Data Mining: Data Mining Concepts and Techniques, International Conference on Machine Intelligence Research and Advancement, 2013.

[45] Ferreira et al. Applying data mining techniques to improve diagnosis in neonatal jaundice, BMC Medical Informatics and Decision Making, 2012.

[46] Sankaranarayanan.S, et al, Diabetic prognosis through Data Mining Methods and Techniques, International Conference on Intelligent Computing Applications, 2014.

[47] B.Venkatalakshmi, et al, Heart Disease Diagnosis Using Predictive Data mining, International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014.

[48] Chao-Ton Su, et al, Data Mining Techniques for Assisting the Diagnosis of Pressure Ulcer Development in Surgical Patients, J Med Syst, 2012.

[49] V. Krishnaiah et al, Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013, 39 – 45.

[50] Sankar K. Pal and Simon C. K. Shiu, Foundations of Soft Case-Based Reasoning.CH-1, ISBN 0-471-08635-5 Copyright @ 2004 John Wiley & Sons, Inc.

[51] Jim prentzas, et al, Case Based Reasoning Integrations: Approaches and Applications, Nova science publishers 2009.

[52] Negny Ste´phane, et al, Effective retrieval and new indexing method for case based reasoning: Application in chemical process design Engineering, Applications of Artificial Intelligence 23, 2010.

[53] Jyh-Bin Yang & Nie-Jia Yau (2000) Integrating case-based reasoning and expert system techniques for solving experience-oriented problems, Journal of the Chinese Institute of Engineers, 23:1, 83-95,

[54] Shweta Tyagi, et al, A Hybrid Recommender System Using Rule-Based and Case-Based Reasoning, International Journal of Information and Electronics Engineering, Vol. 2, No. 4, July 2012.

[55] K. Ashwin Kumar, et al, Hybrid approach using case-based reasoning and rule-based reasoning for domain independent clinical decision support in ICU, Expert Systems with Applications 36 (2009) 65–71.

[56] M.-J. Huang et al., Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Systems with Applications 32 (2007) 856–867.

[57] Z.Y. Zhuang et al., Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners European Journal of Operational Research 195 (2009) 662–675.

[58] Vimala Balakrishnan, et al, Predictions Using Data Mining and Case-based Reasoning: A Case Study for Retinopathy, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering Vol:6, No:3, 2012.

[59] Douglas Bell, Software Engineering for Students: A Programming Approach, Fourth Edition (Book) 2005.

[60] Roger S. Pressman, Ph.D. Software Engineering: A Practitioner's Approach Seventh Edition 2010.

[61] http://www.cs.waikato.ac.nz/ml/weka/index.html N.D

[62] Gökhan Ozar, MySQL Management and Administration with Navicat, September 2012, Packt Publishing Ltd.

[63] Andreas, JavaFX Programming Cookbook, Exelixis Media P.C., 2016.

[64] Irina Fedortsova, Greg Brown, JavaFX/Mastering FXML, Release 2.2, Oracle, 2014.

[65] John S. Bradley, Carrie L. Byington, et al, (2011) The Management of Community-Acquired Pneumonia in Infants and Children Older than 3 Months of Age: Clinical Practice Guidelines by the Pediatric Infectious Diseases Society and the Infectious Diseases Society of America.

[66] Endale Teferra, Hailubeza Alemu, et al, (2010), Standard Treatment Guideline For General Hospitals, Drug Administration and Control Authority of Ethiopia Contents.

[67] Kongmany Chaleunvong, Data collection techniques, Laos Vientiane, 25 September 2009.

[68] http://health.facty.com/ailments/respiratory/10-cough-home-remedies/4/

[69] Mohammed J. Zaki, Limsoon Wong, Data Mining Techniques, 2003.

[70] Manisha Girotra, Kanika Nagpal, et al, Comparative Survey on Association Rule Mining Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 84 – No 10, December 2013.

[71] McSherry, Precision and Recall in Interactive Case-Based Reasoning, In Case Based Reasoning Research and Development (ICCBR), Lecture notes in Artificial Intelligence, pp. 392-306, 2001.

[72] Henok B, A Case-Based Reasoning Knowledge Based System for Hypertension Management. Unpublished Master's Thesis, Addis Ababa University, Ethiopia, 2011.

[73] Getachew W. (2012).Application of case-based reasoning for anxiety disorder diagnosis, Addis Ababa University, Ethiopia.

[74] Biazen G. (2013). Application of case based recommender system in field of study selection, Addis Ababa University, Ethiopia.

[75] Yibeltal C. (2013), Application of case based recommender system in investment Sector and investment activity selection to new investors: in the case of Ethiopia, Addis Ababa University, Ethiopia.

APPENDIXES

APPENDIX I: PROTOTYPE EVALUATION FORM FOR THE DOMAIN EXPERT

This is an evaluation form to be filled by pneumonia disease diagnosis experts in order to evaluate the applicability of the case-based reasoning system in pneumonia disease diagnosis. I thank you in advance for your willingness and valuable time. Parameters of evaluation criteria for user acceptance test is created from the standard metrics of software quality from user perspectives.

Description of the parameter values are as follows.

Performance value	1	2	3	4	5
Description	Poor	Fair	Good	Very good	Excellent

Instruction: please assign (X) on the appropriate value for the corresponding parameter of evaluation questions of the case based reasoning system in pneumonia disease diagnosis.

NO	Parameters of evaluation criteria	Performance value				
		1	2	3	4	5
1	Easy to use of the case based reasoning system					
2	Is the system efficient in time					
3	Is the user interface interactive					
4	Adequacy and clarity of decision support					
5	Relevancy of the retrieved case in the decision making					
6	Fitness of the final solution to the new case					
7	Relevancy of the attributes in representing patient case					
8	Does the explanation facility give brief description about the recommended health service					
9	Rate the significance of the system in the domain area					

APPENDIX II: SAMPLE PNEUMONIA CASES FROM EXCEL SNAPSHOT

CASE NO	Age	Gende	CoughType	FeverLevel	Appetite	Sweating	ShortenessOft	ChestPain	ChestIndrawin	NauseaVomiting	CompleteBloodCount	Chest-X-Ray	Types of Pneumonia	Prescription
Patient1	65	Female	Productive	Fever	Poor	Present	Absent	Absent	Absent	Absent	Not Checked	Not Checked	Community Acquired Pneumonia	Azijhormycin 500mg,polday,cefroxin 19m IVBID
Patient2	70	Male	Productive	Fever	Good	Present	Absent	Present	Absent	Absent	Not Checked	Not Checked	Streptococcus Community Acquire	Azijhormycin 500mg,polday#3,cefroxin 19m IVBI
Patient3	20	Male	Productive	Fever	Good	Present	Present	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	INO2,cefroxin 19m IVBID
Patient4	75	Male	Productive	Fever	Good	Absent	Present	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	Azijhormycin 500mg,polday,cefroxin 19m IVBID
Patient5	20	Male	Mild Cough	NO Fever	Good	Absent	Absent	Absent	Absent	Absent	Not Checked	Not Checked	Pneumonia	Diclofenac 75mg, IM BID , Heamup Syvup
Patient6	56	Female	Mild Cough	Fever	Good	Absent	Absent	Present	Absent	Absent	Not Checked	Not Checked	Streptococcus Community Acquire	cefroxin 19m IVBID, Azijhormycin 500mg,polday#
Patient7	3	Male	Mild Cough	Fever	Good	Absent	Present	Absent	Present	Absent	Not Checked	Not Checked	Streptococcus Community Acquire	cefroxin 19m IVBID, Agumentin228mg/5ml
Patient8	29	Male	Mild Cough	NO Fever	Good	Absent	Present	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus Community Acquire	Azijhormycin 500mg,polday#3,cefroxin 19m IVBI
Patient9	23	Female	Mild Cough	NO Fever	Good	Absent	Present	Present	Absent	Absent	Not Checked	Not Checked	Streptococcus Community Acquire	Azijhormycin 500mg,polday#3,cefroxin 19m IVBI
Patient10	25	Male	Mild Cough	High Grad	Good	Absent	Absent	Absent	Absent	Absent	Not Checked	Not Checked	Pneumonia	cefroxin 19m IVBID,Klamox 625 mg poBID #7day
Patient11	30	Male	Dry	Fever	Poor	Absent	Present	Absent	Absent	Absent	Not Checked	Not Checked	Atypical pneumonia	ceftroxone 19m IVBID, Azithwmycin 500mg, pold
Patient12	60	Male	Mild Cough	NO Fever	Poor	Absent	Absent	Present	Absent	Absent	Not Checked	Not Checked	Community Acquired Pneumonia	ceftroxone 19m IVBID,ION2
Patient13	22	Male	Mild Cough	Fever	Good	Absent	Absent	Present	Absent	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	cefroxin 19m IVBID, INO2, Azijhormycin 500mg,p
Patient14	25	Female	Mild Cough	NO Fever	Good	Absent	Present	Present	Absent	Absent	Not Checked	Not Checked	Pneumonia	Azijhormycin 500mg,polday #5,cefroxin 19m IVBI
Patient15	80	Male	NO Cough	High Grad	Poor	Absent	Absent	Absent	Absent	Absent	Not Checked	Not Checked	Pneumonia	Azijhormycin 500mg,polday #3, M/F
Patient16	75	Male	Productive	Fever	Poor	Present	Absent	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	Cefroxin 19m IVBID, Diceofene 75mg IM
Patient17	45	Male	NO Cough	Fever	Poor	Absent	Absent	Present	Absent	Absent	Not Checked	Not Checked	Community Acquired Pneumonia	Azijhormycin 500mg,polday ,Cefroxin 19m IVBID
Patient18	42	Male	Mild Cough	NO Fever	Poor	Present	Absent	Present	Absent	Absent	Not Checked	Not Checked	Community Acquired Pneumonia	Cefroxin 19m IVBID, Azijhormycin 500mg,polday
Patient19	80	Male	Mild Cough	Fever	Poor	Absent	Absent	Absent	Absent	Absent	Not Checked	Not Checked	Pneumonia	Azijhormycin 500mg,polday #3 ,Cefroxin 19m IVE
Patient20	1	Male	Mild Cough	Low Grad	Poor	Absent	Absent	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus Community Acquire	INO2,Parcetemol suppose PRN, M/F
Patient21	6MOTH	Male	Mild Cough	Fever	Good	Absent	Present	Present	Absent	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	Crystalin pericilline 637, IV QID, INO2, M/F
Patient22	9MOTH	Female	Mild Cough	High Grad	Poor	Absent	Present	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus Community Acquire	INO2,Crystalin pericilline 680,500, IV QID, Parce
Patient23	3MOTH	Male	Mild Cough	NO Fever	Poor	Absent	Absent	Absent	Present	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	INO2,Crystalin pericilline 750,000, IV QID,CAF 2
Patient24	9MOTH	Male	Mild Cough	Fever	Poor	Absent	Absent	Absent	Absent	Absent	Positive	Positive	Streptococcus Community Acquire	INO2, Cefroxin 19m IV BID
Patient25	5MOTH	Female	Mild Cough	Fever	Poor	Absent	Absent	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	INO2,Crystalin pericilline 375,000, IU IVQID, Par
Patient26	10MOTH	Female	Mild Cough	NO Fever	Poor	Absent	Present	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	Crystalin pericilline 750,000, IU IVQID, INO2, Pa
Patient27	9MONT	Male	Mild Cough	Low Grad	Good	Absent	Present	Absent	Absent	Absent	Not Checked	Not Checked	Streptococcus pneumoniae	Crystalin pericilline 375,000, IU IVQID, Parcetem

APPENDIX III: SAMPLE ASSOCIATION RULES GENERATED USING APRIORI ALGORITHM

1. Appetite=Poor Sweating=Absent ChestIndrawing=Absent 342 ==> NauseaVomitingDiarrhea=Absent 340 <conf:(0.99)> lift:(1.61) lev:(0.13) [129] conv:(43.81)
2. CoughType=Mild Cough ChestIndrawing=Absent NauseaVomitingDiarrhea=Absent 323 ==> Sweating=Absent 321 <conf:(0.99)> lift:(1.48) lev:(0.1) [104] conv:(35.5)
3. Appetite=Poor NauseaVomitingDiarrhea=Absent CompleteBloodCount=Not Checked 367 ==> ChestIndrawing=Absent 359 <conf:(0.98)> lift:(1.24) lev:(0.07) [68] conv:(8.54)
4. CoughType=Mild Cough NauseaVomitingDiarrhea=Absent 363 ==> Sweating=Absent 354 <conf:(0.98)> lift:(1.45) lev:(0.11) [110] conv:(11.97)
5. Appetite=Poor ChestPain=Absent NauseaVomitingDiarrhea=Absent 336 ==> ChestIndrawing=Absent 327 <conf:(0.97)> lift:(1.23) lev:(0.06) [61] conv:(7.04)
6. Appetite=Poor Sweating=Absent NauseaVomitingDiarrhea=Absent 352 ==> ChestIndrawing=Absent 340 <conf:(0.97)> lift:(1.22) lev:(0.06) [61] conv:(5.67)
7. NauseaVomitingDiarrhea=Absent Chest-X-Ray=Not Checked 370 ==> ChestIndrawing=Absent 357 <conf:(0.96)> lift:(1.22) lev:(0.06) [64] conv:(5.54)
8. Appetite=Poor NauseaVomitingDiarrhea=Absent 481 ==> ChestIndrawing=Absent 461 <conf:(0.96)> lift:(1.21) lev:(0.08) [80] conv:(4.8)
9. ChestPain=Absent NauseaVomitingDiarrhea=Absent 383 ==> ChestIndrawing=Absent 367 <conf:(0.96)> lift:(1.21) lev:(0.06) [64] conv:(4.72)
10. Sweating=Absent ChestIndrawing=Absent CompleteBloodCount=Not Checked 352 ==> NauseaVomitingDiarrhea=Absent 337 <conf:(0.96)> lift:(1.55) lev:(0.12) [120] conv:(8.45)
11. Sweating=Absent Chest-X-Ray=Not Checked 363 ==> ChestIndrawing=Absent 346 <conf:(0.95)> lift:(1.21) lev:(0.06) [59] conv:(4.23)
12. NauseaVomitingDiarrhea=Absent CompleteBloodCount=Not Checked 456 ==> ChestIndrawing=Absent 432 <conf:(0.95)> lift:(1.2) lev:(0.07) [71] conv:(3.82)
13. Appetite=Poor ShortenessOfBreath=Absent 359 ==> ChestIndrawing=Absent 338 <conf:(0.94)> lift:(1.19) lev:(0.05) [54] conv:(3.42)

14. Sweating=Absent NauseaVomitingDiarrhea=Absent CompleteBloodCount=Not Checked 358 ==> ChestIndrawing=Absent 337 <conf:(0.94)> lift:(1.19) lev:(0.05) [54] conv:(3.41)
15. Appetite=Poor ChestIndrawing=Absent CompleteBloodCount=Not Checked 383 ==> NauseaVomitingDiarrhea=Absent 359 <conf:(0.94)> lift:(1.52) lev:(0.12) [123] conv:(5.89)
16. Sweating=Absent NauseaVomitingDiarrhea=Absent 481 ==> ChestIndrawing=Absent 447 <conf:(0.93)> lift:(1.18) lev:(0.07) [66] conv:(2.88)
17. ShortenessOfBreath=Absent 416 ==> ChestIndrawing=Absent 386 <conf:(0.93)> lift:(1.17) lev:(0.06) [57] conv:(2.81)
18. NauseaVomitingDiarrhea=Absent 620 ==> ChestIndrawing=Absent 575 <conf:(0.93)> lift:(1.17) lev:(0.08) [84] conv:(2.82)
19. CompleteBloodCount=Not Checked Chest-X-Ray=Not Checked 363 ==> ChestIndrawing=Absent 336 <conf:(0.93)> lift:(1.17) lev:(0.05) [49] conv:(2.72)
20. CoughType=Mild Cough Sweating=Absent ChestIndrawing=Absent 347 ==> NauseaVomitingDiarrhea=Absent 321 <conf:(0.93)> lift:(1.5) lev:(0.11) [107] conv:(4.94)

APPENDIX IV: THE RETAIN DIALOG OF CBRDM SYSTEM