



JIMMA UNIVERSITY
JIMMA INSTITUTE OF TECHNOLOGY
FACULTY OF COMPUTING AND INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE
MSc. IN INFORMATION AND KNOWLEDGE MANAGEMENT

**DESIGNING A CASE BASED REASONING SYSTEM FOR DIAGNOSIS
AND TREATMENT OF PNEUMONIA FOR UNDER-FIVE YEAR
CHILDREN IN CASE OF JIMMA UNIVERSITY SPECIALIZED
HOSPITAL**

BY: DANIEL BEKELE

February, 2021
Jimma, Ethiopia

JIMMA UNIVERSITY
JIMMA INSTITUTE OF TECHNOLOGY
FACULTY OF COMPUTING AND INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE
MSc. IN INFORMATION AND KNOWLEDGE MANAGEMENT

**DESIGNING A CASE BASED REASONING SYSTEM FOR DIAGNOSIS
AND TREATMENT OF PNEUMONIA FOR UNDER-FIVE YEAR
CHILDREN IN CASE OF JIMMA UNIVERSITY SPECIALIZED
HOSPITAL**

**A Thesis Submitted in Partial Fulfillment of the Requirements for
Degree of Masters of Science in Information Science
(Information and Knowledge Management)**

BY:

DANIEL BEKELE

Principal advisor: Amanuel Ayde (Assis.Prof. and PhD. Candidate)

Co-Advisor: Muktar Bedaso (MSc.)

February, 2021
Jimma, Ethiopia

JIMMA UNIVERSITY
JIMMA INSTITUTE OF TECHNOLOGY
FACULTY OF COMPUTING AND INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE
MSc. IN INFORMATION AND KNOWLEDGE MANAGEMENT

**DESIGNING A CASE BASED REASONING SYSTEM FOR DIAGNOSIS
AND TREATMENT OF PNEUMONIA FOR UNDER-FIVE YEAR
CHILDREN IN CASE OF JIMMA UNIVERSITY SPECIALIZED
HOSPITAL**

BY:

DANIEL BEKELE

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
Amanuel Ayde (Assis.Prof. and PhD. Candidate)	Principal Advisor	_____	_____
Muktar Bedaso (MSc.)	Co-Advisor	_____	_____
_____	External Examiner	_____	_____
_____	Internal Examiner	_____	_____

DECLARATION

I declare that this thesis is my original work and it has not been presented for a degree in any other university. All the material sources used in this work are appropriately acknowledged.

Daniel Bekele

February, 2021

This thesis has been submitted to the department for examination with our approval as university advisors:

	Name	Sign	Signature
Principal Advisor:	Amanuel Ayde (Ass. Prof, PhD candidate.)	_____	_____
Co-Advisor:	Muktar Bedaso (MSc.)	_____	_____

February, 2021

DEDICATION

This work is dedicated to my beloved families.

ACKNOWLEDGEMENT

First of all, I would like to gratitude almighty God for giving me strength and wisdom to complete this thesis work. Next to God, I would like to take this opportunity to express my profound gratitude and deep regard to my advisors, Amanuel Ayde and Muktar Bedaso, for their exemplary guidance, valuable feedback and constant encouragement throughout the duration of the research. Their valuable suggestions were of immense help throughout the research.

I express my warm thanks to Jimma University Department of Information Science staff, my families and all the friends and colleagues who provided me with the facilities being required and conducive conditions for this research. Especially, I would like to express my heartfelt thanks to my friends Daniel Getachew and Betsegaw Desalegn for their day to day directions and suggestions. I would also like to thank Jimma University Specialized Hospital Pediatrics department and patient record office staff for their kindly help and support.

Finally, I express my warm thanks to my families and to all my friends for their moral support and encouragements throughout my study.

Table of Contents

List of Tables	vi
List of Figures	vii
List of Abbreviations and Acronyms	viii
ABSTRACT	ix
CHAPTER ONE	1
INTRODUCTION	1
1.1. Background of the Study.....	1
1.2. Statement of the Problem	4
1.3. Research Questions	7
1.4. Research Objectives.....	7
1.5.1. General Objective	7
1.5.2. Specific Objective.....	8
1.5. Scope and Limitation of the Study.....	8
1.6. Significance of the Study.....	9
1.7. Organization of the Study	9
CHAPTER TWO	11
LITERATURE REVIEW AND RELATED WORK	11
2.1. Artificial Intelligence	11
2.2. Knowledge Based System	12
2.3. Knowledge Base System Development.....	12
2.3.1. Knowledge Acquisition.....	13
2.3.2. Knowledge Modeling.....	13
2.3.3. Knowledge Representation	14
2.4. Case-Based Reasoning Cycle.....	15
2.4.1. Retrieve	16
2.4.2. Reuse.....	17
2.4.3. Revise.....	17
2.4.4. Retain	17
2.5. Case-Based System Evaluation Method.....	18
2.6. Case-Based System Development Tools	19
2.7. Data Mining.....	21

2.7.1. Data Mining Process Models	22
2.7.1.1. The KDD Process Model	22
2.7.1.2. CRISP-Data Mining Process Model	22
2.7.2. Data mining Classification Techniques and Algorithms.....	24
2.7.2.1. Decision Tree	25
2.7.2.2. Rule Based Classification	26
2.7.2.3. Bayesian Network Classifiers	27
2.8. Related Works.....	28
CHAPTER THREE.....	36
METHODOLOGY	36
3. Methodology of the Study	36
3.1. Problem identification and Motivation	37
3.2. Objectives of the solution	39
3.3. Design and Development Approaches for Knowledge Use.....	39
3.4. Demonstration.....	41
3.5. Evaluation Methods	42
CHAPTER FOUR.....	43
KNOWLEDGE ACQUISITION, MODELING AND EXPERIMENTATION.....	43
4. Knowledge Acquisition.....	43
4.1. Manual Knowledge Acquisition	43
4.1.1. Treatment and Types of Pneumonia	45
4.1.3. Pneumonia Category	55
4.2. Case Modeling.....	56
4.2.1. Conceptual Modeling Using Decision Tree.....	57
4.3. Knowledge Acquired from Data Mining.....	60
4.3.1. Data preprocessing.....	61
4.3.1.1. Data cleaning.....	62
4.3.1.2. Attribute selection.....	63
4.3.1.3. Data Transformation	66
4.3.1.4. Data Formatting	67
4.4. Experimentation	68
4.4.1 Experiment Design.....	68

4.4.2 Comparison of classification algorithms.....	78
CHAPTER FIVE	84
SYSTEM DEVELOPMENT AND EVALUATION	84
5.1. Architecture of the Prototype System	84
5.2. Case Based Reasoning System for CBRSDTPUFYC	86
5.2.1. Building a Case-Base	87
5.2.2. Case Representation.....	87
5.2.3. Managing the Case Structure in JCOLIBRI.....	87
5.2.4. Description of CBRSDTPUFYC Case Attributes.....	88
5.2.5. Managing Connectors	90
5.2.6. Managing Tasks and Methods	91
5.3. System Evaluation.....	96
5.3.1. Testing the CBR Cycles and Evaluating the Performance of the prototype.....	96
5.3.1.1. Retrieve and Reuse Evaluation	96
5.3.2.2. Case Revision and Solution Adaptation Testing.....	96
5.3.2. User Acceptance Testing.....	100
5.4. Discussion of Results.....	101
CHAPTER SIX CONCLUSION AND RECOMMENDATIONS.....	106
6.1. Conclusion.....	106
6.2. Recommendations	107
References	109
APPENDICES I	118
APPENDIX I: INTERVIEW QUESTIONS.....	118
APPENDIX II: USER ACCEPTANCE TEST.....	119
APPENDIX III: PART CLASSIFIER OUTPUTS.....	120

List of Tables

Table 4.1:- Bacterial pneumonia treatment.....	45
Table 4.2:- Viral pneumonia treatment.....	47
Table 4.3:- Aspiration pneumonia treatment	48
Table 4.4:- Mycoplasma pneumonia treatment	50
Table 4.5:- Fungal pneumonia treatment	52
Table 4.6:- Broncho pneumonia treatment	53
Table 4 7:- Ventilator-associated pneumonia treatment.....	54
Table 4.8:- Streptococcus pneumonia treatment.....	55
Table 4.9:- Removed attributes.....	62
Table 4.10:-Selected attributes with their description	63
Table 4.11:-Discretized attributes with values.....	67
Table 4.12:-Sample ARFF used for classification.....	68
Table 4.13:-Attribute name used for experimentation analysis	69
Table 4.14:-Confusion matrix of J48 decision tree with 10-fold cross validation.....	69
Table 4.15:-Summary of J48 decision tree classifier experiment result.....	70
Table 4.16:-Confusion matrix of J48 decision tree with percentage split	71
Table 4.17:-Summary of J48 decision tree classifier experiment result.....	72
Table 4.18:-Confusion matrix of PART with 10-fold cross validation	73
Table 4.19:- Summary of PART classifier experiment result for 10-fold cross validation.....	74
Table 4.20:- Summary of PART classifier experiment result with percentage split	75
Table 4.21:- Summary of PART classifier experiment result.....	75
Table 4.22:-Confusion matrix of Naïve Bayes with 10-fold cross validation	76
Table 4.23:-Summary of Naïve Bayes classifier experiment result	76
Table 4.24:-Confusion matrix of Naïve Bayes with percentage splitter.....	77
Table 4.25:- Summary of Naïve Bayes classifier experiment result	77
Table 4.26:-Comparison of classification algorithms.....	78
Table 5.1:- List of attributes and description	89
Table 5.2:-Relevant cases assigned by domain experts for sample test cases	99
Table 5.3:-User Acceptance testing from domain experts.....	100

Table 5.4:-Comparison of the developed CBR prototype system with previous studies	102
---	-----

List of Figures

Figure 2 1:- Development of a Knowledge-Based System.....	13
Figure 2 2:- CBR cycle	15
Figure 3 1:- Design Science research process model adopted from	37
Figure 4 1:- Decision Tree for Diagnosis and Treatment of Pneumonia	60
Figure 4 2:-Information Gain result for attribute selection.....	66
Figure 4 4:-A tree generated from J48 pruned tree	73
Figure 4 5:-A sample of test instances using PART classifier algorithm	80
Figure 4 6:-Model building result through commands on WEKA “Simple CLI”	81
Figure 4 7:-Sample CLI prediction results on test data using PART	83
Figure 5.1:-Architecture of the CBRSDTPUFYC system.....	86
Figure 5.2:- Main and Configuration Window of JCOLIBRI	86
Figure 5.3:- Managing Case Structure In jCOLIBRI.....	88
Figure 5.4:- JCOLIBRI Connector Schema.....	90
Figure 5.5:- Managing Connectors window	91
Figure 5.6:- Window for Obtaining query task.....	92
Figure 5.7:- Window of Case Revision.....	93
Figure 5.8:-Tasks and methods configuration	95
Figure 5.9:-Query Interface	97
Figure 5.10:-Retrieved Solution case.....	98
Figure 5.11:-Revision Interface	98
Figure 5.12:-Retaining the revised case.....	99
Figure 5.13:-Interface of learned cases	99

List of Abbreviations and Acronyms

AI:	Artificial Intelligence.
AIDS:	Acquired Immune Deficiency Syndrome
CBR:	Case Based Reasoning
CBRSDTPUFYC	Case Based Reasoning System Diagnosis and Treatment of Pneumonia Under-Five Year Children
CRISP-DM	CRoss-Industry Standard Process for Data Mining
DM:	Data Mining.
HP	Hewlett-Packard
GAIA	Group for Artificial Intelligence Applications
GUI:	Graphical User Interface
IDE:	Integrated Development environment
JCOLIBRI	Java Case and Ontology Libraries Integration for Building Reasoning infrastructures
JDK:	Java Development Kit
KBS:	Knowledge-Based System
KDD	Knowledge Discover Database
MAC	Media Access Control
MS-DOS	Microsoft disk operating system
RDBMS:	Relational Database Management System
RBR:	Rule-Based Reasoning
WHO	World Health Organization
WEKA	Waikato Environment for Knowledge Analysis
XML	Extensible Markup Language

ABSTRACT

Pneumonia is the single leading cause of mortality in under five year children and is a major cause of child mortality in every region of the world, with most deaths occurring in sub Saharan Africa and South Asia. It is also known to be one of the predominant causes of mortality for under-five children in Ethiopia and lack of sufficient pediatricians. Since, conducting this study is very important to minimize death rate. The main objective of this study is to develop a case-based system for the diagnosing and treatment of pneumonia under five-year children. The study employed a design science research approach to understand the problems in the area and develop model. The researcher used manual and automated knowledge acquisition techniques, such as interview, document analysis and data mining. For this study, predictive data mining task mainly classification technique was performed to generate representative cases from the prepared data. The required data were acquired from Jimma University Specialized Hospital. WEKA data mining tool is used for experimentation. Three experiments were conducted by using J48, PART, and Naïve Bayes classification algorithms to identify the best model and select the best performing data mining classification algorithm. Based on experimental result, PART classification algorithm is selected to construct cases for the case based system because it registered better performance than other classifiers. The developed model was tested with test instances and only those instances registers more than 99% accuracy were used to develop a knowledge base for the CBR development for a better efficiency. Then, implement the prototype by using jCOLIBRI version 1.1. Finally, testing of the developed prototype CBR system is done to evaluate the performance of the system. The prototype is evaluated using system testing and user acceptance testing. System testing performed in terms of recall, precision and F-measure registered 96%, 89% and 92.36%, respectively. User acceptance testing also performed by involving domain experts and an average of 94% acceptance was achieved. This shows the system has registered a promising result. However, case-based reasoning system needs to be supported by rule-based reasoning for providing a complete advice for the problem, increasing number of cases and including other significant attributes improve the performance of the developed system which is forwarded as future work.

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

Pneumonia is an infection that inflames the air sacs in one or both lungs. The air sacs may fill with fluid or pus (purulent material), causing cough with phlegm or pus, fever, chills, and difficulty breathing. A variety of organisms, including bacteria, viruses and fungi, can cause pneumonia. Pneumonia is the single leading cause of mortality in children under five and is a major cause of child mortality in every region of the world, with most deaths occurring in sub Saharan Africa and South Asia (WHO, 2020). Pneumonia kills more children under five than AIDS, malaria, and measles combined, yet increased attention in recent years have been on the latter diseases. Pneumonia is a form of acute respiratory tract infection (ARTI) that affects the lungs. When an individual has pneumonia, the alveoli in the lungs are filled with pus and fluid, which makes breathing painful and limits oxygen intake. In order to prevent pneumonia in children is an essential component of a strategy to reduce child mortality. Immunization against Hib (Homophiles influenza type b), pneumococcus, measles and whooping cough (pertussis) is the most effective way to prevent pneumonia (WHO, 2020).

Nowadays, there is an increasing appreciation of the role that computers are playing in improving the overall health delivery system. Specifically, the application of knowledge-based systems is one of the mechanisms that improve health service quality. The concept of knowledge-based systems is derived from the field of artificial intelligence (AI). AI intends understanding of human intelligence and the building of computer programs that are capable of simulating or acting one or more intelligent behaviors. (Priti S & Rajendra A, 2010).

Knowledge-based systems (KBSs) in medicine have received attention, because of the potential benefits that can be gained from using them. They may facilitate increasing productivity in a medical environment, support the making of diagnoses and other types of medical decisions, assist in the training of medical professionals, and can even handle some routine tasks in a medical environment (Abdel and AinShams, 2016). There are different types of case representation techniques used for diagnosis; among which the most common are rule based and case based reasoning (Pandey and Mishra, 2009). Rules represent general knowledge of the

domain, whereas cases represent specific knowledge. Rule-based systems solve problems from scratch, while case-based systems use pre-stored situations to deal with similar new instances. In rule based updatability of solution is challenging and prepared to work only on the existing rules. While CBR is used to reduce the knowledge acquisition task, providing flexibility in knowledge modeling. This indicated that a case based system is an easy technique for developing a knowledge-based system for diagnosis and treatment of pneumonia for under-five year children. In addition, knowledge in the form of new cases faced during real-time operation can be incorporated into the case base in extending the effectiveness of the case based system.

According to (Rainer & Lothar, 2000), Case-based Reasoning (CBR) has become a successful technique for knowledge-based systems in many domains; while in medical domains, some more problems arise with using this method. CBR for medical knowledge-based systems, points out problems, limitations and possibilities how they can partly be overcome. Case-based Reasoning means using previous experience in the form of cases to understand and solve new problems. A case-based reasoning remembers former cases similar to the current problem and attempts to modify their solutions to fit the current case. The underlying idea is the assumption that similar problems have similar solutions (Rainer & Lothar, 2000).

Case based reasoning methodology provides a foundation for a new technology of building intelligent computer aided diagnoses systems (Abdel-Badeeh, 2007). CBR in medicine helps to diagnose and treatment of diseases by using previously successfully solved experiences of specialized doctors. It doesn't mean the CBR approach replaces the work of a specialist doctor but helps in decision making to apply the experience of highly qualified health professionals in their absence (Mekedes, 2018).

CBR is a type of knowledge representation which uses previous experiences in the form of cases to understand and solve new problems. Solving a problem by CBR involves gaining a problem description, measuring the similarity of the current problem to previous problems stored in a case base with their known solutions, retrieving one or more similar cases, and attempting to reuse the solution of one of the retrieved cases, possibly after adapting it to account for differences in problem descriptions. The system's suggested solution is then tested. If we have a new problem, we must represent it as a case, and then we must go through four steps in CBR: retrieve the most

similar case or cases, reuse the case or cases to try to solve the problem, revise the proposed solution if possible, and retain the new solution as part of a new case (Pascal, R et al, 2017).

During the development of KBS, solved knowledge must be acquired because the most important part of any knowledge-based system is the knowledge and the power of an expert system resides in the specific, high-quality knowledge it contains about task domains. Knowledge is incomplete and dynamic. Will extend our options through which we can acquire knowledge from different sources such that we can make the knowledge base of the Knowledge Based System as complete as possible. In order to make knowledge extraction as correct as possible, different techniques could be applied. Among these techniques, data mining or knowledge discovery techniques have become the most used ones in recent years (Mihaela, 2006).

Data mining improves decision making by giving insight into what is happening today and by helping predict what will happen tomorrow (Amritpal et al, 2015). Data mining (DM) is a subfield of Machine Learning that enables finding interesting knowledge (patterns, models and relationships) in very large databases. It is the most essential part of the knowledge-discovery process, which combines databases, statistics, artificial intelligence and machine learning techniques (Mihaela, 2006). Data mining is the extraction of hidden Knowledge from large databases, which is a powerful new technology with great potential to help user focus on the most important information generated from their large data set (Sudhir & Kodge, 2013).

Data mining in health care industry today extract useful knowledge from large amount of complex data for diagnosis and treatment of various diseases, symptom analysis and disease classification or edification purposes. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making (Durairaj & Ranjani, 2013). Data mining is also the unified name for all tools that can be used when searching for relationships and trends in large amounts of data, mainly used on data showing no such trends when judged by the human eye.

The purpose of this study is to develop a case based system for diagnosing and treatment of pneumonia under five-year children by using data mining technique for case acquisition. This study is important because pneumonia is a prevalent disease and the cause of many children's

deaths. Since CBR is similar to the way physicians make reasoning about patients and have the concept of how they use their experience. Patient records collected by hospitals and doctors, can easily be integrated and used with CBR and provide solutions for decision support systems to solve current problems based on similar past solutions. This study will benefit the country and society by supporting activities which are useful in reducing pneumonia in under five years due to the shortcomings of medical experts in the area.

1.2. Statement of the Problem

Pneumonia is the number one infectious killer of children under age 5. Globally, killing an estimated 1.4 million children under the age of five years, accounting for 18% of all deaths of children under five years old worldwide. Pneumonia affects children and families everywhere, but is most prevalent in South Asia and sub-Saharan Africa. Children can be protected from pneumonia, it can be prevented with simple interventions, and treated with low-cost, low-tech medication and care (WHO, 2020; UNICEF, 2019). Ethiopia is among 15 top under five pneumonia high burden countries. In Ethiopia, Pneumonia is the leading cause of morbidity and mortality under-five year children, with an approximately 3,370,000 children experiencing pneumonia every year that attributes to 20% of all causes of deaths and killing more than 40,000 under five-children every year (Zewudu, M et al, 2020).

CBR is driven by two motivations. The first one is the desire to model human behavior (from cognitive science). The second one is the pragmatic desire to develop technology/technique to make AI systems more effective (Leake, 1996). CBR for health science is today both a recognized and well established method (Shahina, 2011). According to Huang, Chen, & Lee , (2007) CBR is an appropriate reasoning method in medicine for some important reasons; cognitive adequateness, explicit experience, duality of objective and subjective knowledge as well as acquisition of subjective knowledge from the new instance problem.

Children's hospitals across the country continue to experience significant shortages in pediatric specialties. Pediatric specialty shortages affect children and their families' ability to receive timely and appropriate care. Despite advances in public health systems in Ethiopia through global partnerships, there is still lack of well-organized pediatric emergency units (Gemechu, J et al, 2018). And also experienced health professionals are not equally distributed in the country for

better nearby diagnosis and treatment (WHO, 2013). In order to solve this problem, there is a need to apply knowledge base systems as a powerful tool with extensive potential in pneumonia.

Health care is highly complex and interdependent system in nature. In diagnose and treatment of pneumonia under five-year children; effective patient care depends on the interaction of emergency physicians, pediatrician, emergency nurses, laboratory and diagnostic x-ray services. If any one of these interdependent components was performed poorly or overwhelmed, delivery of care in the Pediatrics Department will suffer. Thus, the pediatrics department may experience “operational inefficiencies” as a result of inadequate staffing levels, or poor communication with laboratory, x-ray services and shortages in pediatric specialty affect children and their families’ ability to receive timely and appropriate care; and lack of radiologist to read test results correctly are other headaches to diagnosis and treatment of pneumonia under five year children. As a result the process of diagnosing patient face serious problems with the service provided in the hospital. Therefore, designing an efficient knowledge based system would help the institution to minimize the above interrelated problems that is caused due to lack of potential human expert or lack of potential skills to solve the patient's health problem.

Health professionals need updated health information from credible sources to improve their knowledge and provide evidence based health care services to their clients. Most of health professional are working simply by referring to their handout and remembering their school trainings. To fill the knowledge gap between the specificity of single cases. CBR are appropriate for medical knowledge-based systems, point out problems, limitations and possibilities how they can partly be overcome. Therefore as the assumption of the researcher if this system developed it can solve these problems and help nurses at the place where there is lack of pediatrics and the easily be informed in short time and can provide correct treatment.

Medical equipment is an essential health intervention tool used by health professionals for prevention, diagnosis and treatment of disease and for rehabilitation of patients. However, the shortage of medical equipment, either due to unavailability or non-functionality is a barrier to the ability of the health system to deliver quality health services. The World Health Organization estimates that between 50 to 80 percent of medical equipment in developing countries are not functioning and those countries lack technology assessment systems and regulatory controls to prevent importation of inferior medical equipment.

These days there are different studies attempting to design KBS which have been done in the medical domain. Amelework (2017), investigated the applicability of a case-based reasoning approach to developing a knowledge-based system for tuberculosis diagnosis, Zhenjia, Liangping and Runfeng(2020) was conduct a research for comparison and validation of different deep learning model in order to diagnosis of Pneumonia, Melquiades and Haile (2019) was conducted a research to localized knowledge based system for diagnosis and treatment of pediatric Pneumonia. Aiyasha et al (2019) also conducted a study for differentiate the diagnosis of Tuberculosis and Pneumonia using machine learning algorithms. Ermiyas and Hailemicheal (2020) investigated the applicability of a case-based reasoning approach to developing a knowledge-based system for chronic kidney disease diagnosis. Lucky at al. (2017) Conducted research by using a Case-based reasoning approach for diagnosis of Bowel Disease. Hindayati et al. (2020) conducted Diet Calorie Determination System by using Case-Based Reasoning, Mekedes (2018) Conducted research by using a Case-based reasoning approach for diagnosis malnutrition only under five year children and Ababayehu (2015), was developed a user friendly CBR system for diagnosis and treatment of bacterial pneumonia and viral pneumonia diseases. But the researchers used manual knowledge acquisition technique and investigated a pure case based reasoning. Therefore, (Mekedes, 2018; Amelework, 2017; Ermiyas,B & Hailemichael, K, 2019; Ababayehu, 2015)recommended further investigation is needed in different medical domain since it is an active area. Thus, it should be investigated by case-based reasoning as domain experts use their experience in diagnosis and treatment of the diseases to solve the shortage of domain experts and also use automated knowledge acquisition techniques. Furthermore, Kedier (2018) and Desalegn (2017) used data mining techniques to acquire cases for diagnosis and treatment diabetics and used to determine the choice of contraceptive methods health problems. And also Bezahegn (2017) developing a predictive model for pre- diabetes screening by using data mining technology However, acquiring cases through using the data mining default experimentation settings had its own limitation means test cases through supplied test set shows the given cases are classified correctly or incorrectly. The developed prototype system shows an encouraging result as compared with previous studies.

As indicated by literatures from the previous research gap, further study is needed in different medical domains by using data mining technique as data acquisition for KBS development.

Other alternatives like artificial neural networks require processors with parallel processing power, by their structure. As a result, the equipment's realization is dependent. Unexplained functioning of the network: This is the most important problem of ANN. So based on the gaps indicated, this study attempts to develop a prototype CBR system that can use previous patient history of Jimma University Specialized Hospital pneumonia under five-year children to diagnose and treatment of new patients.

Therefore, this study is conducted with the aim of filling the gaps which are stated in the above section. The case-based reasoning system for pneumonia treatment and diagnosis is designed by using data mining technique as knowledge acquisition highly improve the result of the study. As knowledge-based systems are useful when there is a shortage of experts, and when intelligent assistance or training are required for decision making (Priti S & Rajendra A, 2010). So that, this study attempted to develop a CBR that provides the necessary advice for experts so as to enable them to make the necessary diagnosis and treatment of pneumonia under five-year children.

1.3. Research Questions

To the end, this study attempt to explore and answers the following research questions.

- What kind of domain knowledge and attributes are used for the diagnosis and treatment of pneumonia diseases?
- What is the procedure to acquire suitable cases by using DM techniques that can be used by CBR for pneumonia diagnosis and treatment?
- How the acquired cases will be modeled and represented in developing the CBR system?
- How the prototype case-based system works for diagnosis and treatment of pneumonia under five year children?

1.4. Research Objectives

1.5.1. General Objective

The main objective of this study is to develop a case-based system for the diagnosing and treatment of pneumonia under five-year children by using data mining technique.

1.5.2. Specific Objective

To achieve the above general objective of the study, the researcher set the following specific objectives.

- To acquire domain knowledge from previously solved cases, relevant documents and domain expertise.
- To identify attributes and explore suitable data mining classification algorithm for the prediction of pneumonia treatment and diagnosis.
- To model and represent the acquired knowledge from knowledge experts and solved cases.
- To develop a CBR system that provides advice for diagnosis of pneumonia under five-year children.
- To evaluate the performance and user acceptance of the proposed CBR system.

1.5. Scope and Limitation of the Study

The scope of this research is to develop a prototype CBR system by using data mining techniques that provides expert advice for diagnosis and treatment of pneumonia under five-year children. The knowledge is represented by using a case-based reasoning approach in this research. For the development of the CBR prototype system, the researchers used jCOLIBERI and data mining tool WEKA as knowledge acquisition. There are more than 30 different causes of pneumonia, and they are grouped by the cause. The main types of pneumonia that commonly affect under five years children are bacterial, viral, fungi, aspiration and mycoplasma pneumonia (WHO, 2020) are considered in this study. Due to the current essence of the diseases for children under five years, this study does not include other age groups of patients. The researcher acquired domain knowledge from Jimma University specialized hospital from individual patients' card history or cases, doctors, nurses and health care service providers moreover, from books and documents.

This study is intended to design KBS which includes the tasks of knowledge acquisition, modeling, and knowledge representation and develop KBS that provides the necessary expert advice on pneumonia diagnosis and treatment for under five-year children.

There were some challenges that I faced while doing this research. One of the challenges was the difficulty of acquiring more cases during case collection. The reasons for this were: patient cards were incomplete and most patient records were recorded combined with other disease types and all age group patient cards. So it is difficult to identify only pneumonia under five-year children cases from others. Furthermore, COVID-19 was a big challenge for collecting more cases from the hospital. In order to deal with this challenge, the researcher closely worked with domain experts to select the main contributing features of pneumonia for under five-year children. Additionally, this study is limited to using a selected data mining algorithm for learning classification model.

1.6. Significance of the Study

Knowledge-based systems try to solve problem as a human expert like fashion by using knowledge of application (expert) and problem-solving technique. Thus, the immediate beneficiaries of the system are primary health care workers and health professionals working in the diagnosis and treatment of pneumonia under five year disease. And also to reduce the problem of the limited numbers of expert in giving preliminary diagnosis and treatment of pneumonia especially in remote areas of Ethiopia. KBS has its own role by filling the gap, assist medical personnel in the tedious and complication task of diagnosing, when there are shortages of doctors, thereby, offering primary health care for the people. And also used by physicians as a knowledge sharing tool or organizational memory for stakeholders especially for hospitals, clinics and healthcare service centers which have a shortage of professionals and reduce the workload of pediatricians. In addition, the system provides treatment advice effectively and efficiently based on diagnosis result. The CBR system for diagnosis and treatment of pneumonia under-five year children by using data mining techniques and it used to support for decision making accurately and timely. Furthermore, it can help as a benchmark for researchers who are interested in the area to further related research. Finally, the developed prototype system indicates policy makers, strategies or administrative organization for looking alternative solution for filling the shortage of medical experts in the area.

1.7. Organization of the Study

This thesis is organized into six chapters, Chapter one discusses background of the study, Statement of the problem and research questions, objective of the study, scope and limitation of the study and significant of the study. Chapter two discusses about conceptual and related works review that are relevant for this study. In this chapter, the researcher discuss about Case Based System including CBR cycles, knowledge acquisition, knowledge representation, Knowledge Base System architecture, CBR System Performance Evaluation Methods, Knowledge Base Development tools and data mining and its tasks which are relevant for this study. In chapter three, research methodology and approaches discussed.

Chapter four presents the knowledge acquisition process. The focus here is on manual (domain expert interview and documents analysis) and automated knowledge acquisition techniques through data mining. After the manual knowledge acquisition step, the researcher proposed the conceptual model for pneumonia diagnosis and treatment decision making. The researcher presented the Knowledge discovery steps such as data set preparation, preprocessing, predictive model creation and experimentation. The researcher also discussed the results of WEKA classifier algorithms by comparing one to another and selecting the best performing algorithm. The fifth chapter deals with the development and evaluation of the prototype system and discussion of results. The architecture of the new prototype CBRSDTUFYC for decision making is developed. The performance of the prototype is evaluated both the performance of the system and the acceptance of the system by the users.

Finally, in chapter six presents the conclusions and recommendations of the study. In this chapter based on the result obtained from this study the researcher concluded and give recommendation for future work

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

The ambition for computer systems being able to support human experts during complex problem-solving task is a usual topic of AI research. In order to enable a computer system to give rational support when solving problems in a complex application domain, it is necessary to provide it with specific knowledge within that domain. A number of methodologies to realize such knowledge-based systems have been developed, such as, rule-based approach. In recent years, CBR has become a very popular technique for developing knowledge-based systems that can support using specific knowledge. In order to have deep understanding on the problem of this study, it is very important to review several literatures that have been conducted in the field so far. For this reason, related literature such as books, journal articles, proceeding papers, conference papers and manuals some other sources that are retrieved from the internet have been consulted. In addition to this, the researcher also reviewed related works to identify the gap and formulate the problem and research questions of the study so as to understand the domain knowledge, concepts, principles and methods that are important for achieving the research objective.

2.1. Artificial Intelligence

Artificial intelligence (AI) is the intelligence of a machine. Basically, AI is the branch of science to make the machine as intelligent as a human beings for a particular domain (Poonam, T et al, 2011). AI sense machines will improve human abilities in numerous zones. As it is claimed that artificial intelligence is applied widely in the research of computer science and operational research areas. Intelligence is commonly considered as the ability to acquire and apply different skills and knowledge to solve a given problem. In addition, intelligence is also concerned with the use of general mental capability to solve, reason, and learning various situations. In the near Future, intelligent machines will replace human capabilities in many areas (Rupali, k & Deepali, S, 2018).

Artificial Intelligence is playing an important role in understanding and performing intelligent tasks such as reasoning, learning new skills and adapting to new situations and problems. AI applications in healthcare and pharmaceuticals can help detect health conditions early,

deliver preventative services, optimize clinical decision making, and discover new treatments and medications. They can facilitate personalized healthcare and precision medicine, while powering self-monitoring tools, applications and trackers. AI in healthcare offers potential benefits for quality and cost of care (Angel, 2019).

2.2. Knowledge Based System

The concept of KBS is derived from the field of AI. AI is a machine learning (ML) that intends the understanding of human intelligence and building of computer programs that are capable of simulating or acting one or more of intelligent behavior. Intelligence is the capability of observing, learning, remembering, and reasoning (Abeba, 2014). Using Artificial Intelligence (AI) techniques, computers are able to give the diagnosis of a specific disease called as medical expert systems. However, the practical benefits of such automated reasoning systems have fallen short to give independent expert advice about the particular disease.

The purpose of the knowledge-based system is to act as a decision support system or as a second opinion for the doctors in critical cases. Knowledge-based system is developed to incorporate medical knowledge and reasoning strategies into the automation of medical diagnosis. With the help of new approaches in AI which have recently emerged, may overcome some of the limitations inherent in earlier attempts to automate the medical diagnosis system. It is possible to prepare a knowledge-based system for medical diagnosis to assist the junior doctors or doctors who are practicing at remote places. The medical knowledge of a specialized doctor is required for the development of an expert system (Gulavani & Kulkarni, 2009).

2.3. Knowledge Base System Development

The development of KBS is the integration of many components. Figure 2 below shows the overview of knowledge-based system development process (Priti S & Rajendra A, 2010)

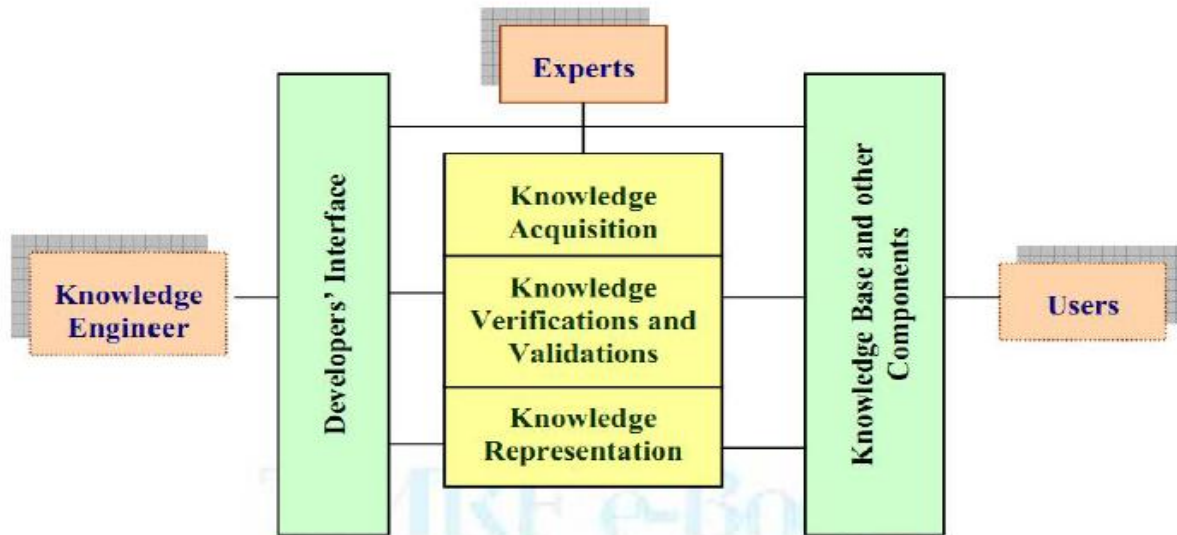


Figure 2 1:- Development of a Knowledge-Based System (*Priti S & Rajendra A, 2010*)

2.3.1. Knowledge Acquisition

Knowledge acquisition (KA) is an important part of developing a KBS using the appropriate methods that should be used for acquiring the knowledge needed for creating and testing the CBR system. Knowledge acquisition is the process of acquiring relevant knowledge from the domain expert, books, documents, sensors, or computer files and structuring and organizing that knowledge into suitable form for knowledge representation. The knowledge can be specific to the problem domain or to the problem-solving procedures, it can be general knowledge (e.g., knowledge about business) or it can be Met knowledge (knowledge about knowledge). Knowledge acquisition is the bottleneck in knowledge-based system development today. Because, the trustworthiness and the performance of the knowledge-based system mainly depend upon the acquired knowledge (Tagel, 2013). The knowledge acquisition process incorporates different methods such as interviews, patient record reviews, observation or document analysis to acquire factual and explicit knowledge. Therefore, all acquisition process are suitable for this study.

2.3.2. Knowledge Modeling

After the knowledge is acquired from pneumonia cases, books, domain experts (health professionals) and other relevant documents, the next step is modeling the knowledge. The knowledge modeling involves organizing and structuring of the knowledge gathered during

knowledge acquisition. This activity provides an implementation-independent specification of the knowledge to be represented in the knowledge base. Knowledge modeling is the concept of representing information and the logic for purpose of capturing, sharing and processing knowledge to simulate intelligence. Here, the basic concepts that reveal the main activities and decisions that are made to solve cases in the domain are modeled (Henok, 2011).

Knowledge modeling is a crucial step in the knowledge acquisition process so as to understand well, the problem domain and to prepare the knowledge representation phase. There are different conceptual modeling techniques. For this study, the researcher used the decision tree knowledge modeling technique because it is suitable for modeling CBR. Decision trees commonly acts as a key role in the knowledge modeling process (Siraj, 2019).

2.3.3. Knowledge Representation

In the Previous section, knowledge has been acquired and modeled; the next step is knowledge representation by using a suitable format that is understandable by the inference engine. Knowledge representation is a means of encoding the human expert knowledge in an appropriate way. The two most known approaches for problem-solving in intelligent systems are case-based and Rule-based reasoning. One of the main differences between a case-based and rule-based reasoning system is on the method in which knowledge is stored and used (Alec, H and George, B, 1999).

Rules and cases are another ways of representing knowledge of an application domain. Rules are used in rule-based reasoning while cases are used in case-based reasoning. Rules are suitable to represent general and normative knowledge, whereas cases are suitable for detailed specific situations. For this study, the researcher used the Case-based knowledge representation techniques to represent the acquired knowledge. CBR is a type of case representation that uses previous experiences inform of cases to understand and solve new problems, cases are capable of representing specific historical knowledge. Cases are natural and easy to acquire. They can be collected from historical records, repair logs, or other sources; eliminate challenging of knowledge acquisition from experts (Alec, H and George, B, 1999). For this study, the researcher uses case-base knowledge representation method for system development which one

of the most predominant and popular knowledge representation methods in the development of Knowledge based System.

2.4. Case-Based Reasoning Cycle

Case-based Reasoning (CBR) has become a successful technique for knowledge-based systems in many domains CBR provides solutions for decision support systems to solve new problems. These solutions are based on similar past solutions. Each of these past experiences is called a Case. Each case consists explanation of the problem and part of the solution. CBR method has a strong ability to learn, it can learn from the experiences of the past, to deal with new problems. The CBR methodology is based on a cycle, namely, the R4 model, composed by four phases: retrieve, reuse, revise and retain (Pascal, R et al, 2017; Narina T et al, 2016). In the first phase, when a new problem is logged, CBR retrieves the most similar case to the problem. In the second phase, the retrieved solution is reused. In the third phase, the solution will be reviewed to fit the new problem. And in the fourth step, the reviewed solution is maintained and retained for future reuse. Obviously, CBR is best appropriate for knowledge-based decisions (Hamid T et al, 2015).

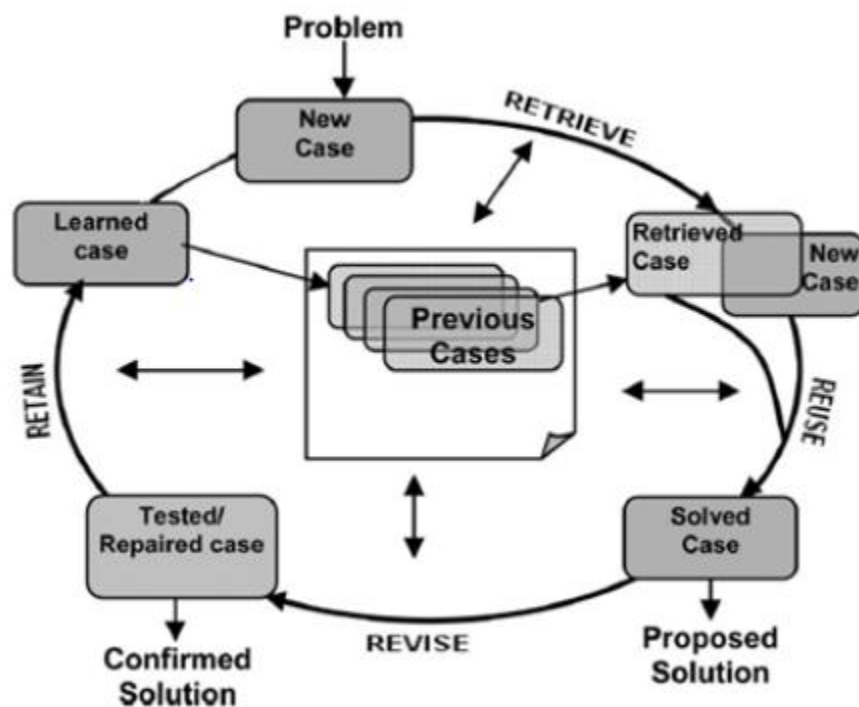


Figure 2 2:- CBR cycle

2.4.1. Retrieve

The most important task of CBR is the retrieval of appropriate cases. Recalling past cases is done based on the similarities between the current case and past cases (Krishnamoorthy and Rajeev, 1996). It takes the description of a problem as its input and provides the best-matched case or set of cases as output. The quality of a case-based reasoning system as a whole is highly affected by the quality of its retrieval process due to its being the base for the rest of the processes.

The quality of the retrieval process depends on its descriptive feature identifying algorithm, searching algorithm and similarity assessment method. The two of most well-known algorithm for case retrieval are: (Singh et al, 2007; Watson & Marir, 1994)

Nearest neighbor algorithm

The Nearest Neighbor algorithm measures the similarity of stored cases with a new input case, based on matching a weighted sum of features (Zhongzhi, 2011). When a new case doesn't exactly match with hold cases then this algorithm will return the nearest match from a case-based reasoning library. It is suitable when there are attributes that have numeric (continuous) value, the algorithm defines and calculates the near value (or the match value) between the cases, and the case with the nearest value is the one that we can use to refer (Fang & Songdong, 2007). But the retrieval time by this algorithm increases linearly as the case in the case base increases.

Induction

Induction is a technique developed by machine learning researchers to extract rules or construct decision trees from past solved cases. In case-based reasoning systems, it analyzes the case base in order to construct a decision tree that classifies the cases. The most popular induction algorithm in case-based reasoning is called ID3. It uses a heuristic called information gain to find the most promising attribute on which to divide the case base (Mohamed et al, 2014)

Induction algorithm is helpful when a single case feature, which is dependent upon others are required as a solution. This algorithm identifies which features do the best job indiscriminating cases and generate a decision tree type structure to organize the cases in memory (Watson & Marir, 1994).

2.4.2. Reuse

After selecting one or several similar cases, the reuse step tries to apply the contained solution information to solve the new problem. Often a direct reuse of a retrieved solution is impossible due to differences between the current and the old problem situation. Then the retrieved solutions have to be modified in order to fit the new situation. How this adaptation is performed strongly depends on the particular application scenario (Wilke, W. & Bergmann, R, 1998).

Proposing a solution can be performed into two ways: reusing the solution as it is or by adapting it. When the selected case and the new case do not have a significant difference, the solution in the selected case will be proposed as it is for the new problem. Whereas, if there is a significant difference between them, the solution in the selected case is adapted based on the unique feature of the new case, this process is known as adaptation (Ramon, L et al, 2006).

2.4.3. Revise

Depending on the employed adaptation procedure, the correctness of the suggested solution often cannot be guaranteed immediately. Then it becomes necessary to revise the solved case. How such a revision is performed, strongly depends on the particular application scenario. For example, it might be possible to apply the suggested solution in the real-world to see whether it works or not. However, often a direct application of an uncertain solution is impossible due to the corresponding risks. Then the revision has to be performed manually by a human domain expert or by alternative methods such as computer simulation (Agnar, A & Enric, P, 1994).

2.4.4. Retain

If the solved case has passed the revising step successfully, a tested/repaired case will be available representing a new experience that might be used to solve similar problems in the future. According to (Smyth & McKenna, 2001) the retain phase is the learning phase of a CBR system. The typical form of learning that occurs in a CBR system is learning by adding a revised case to the case base. Thereby, the new problem-solving experience becomes available for reuse in future problem-solving episodes. The task of the CBR cycle's last step is to retain this new case knowledge for future usage. Therefore, the new case may be added to the case base. In most cases, a general storage of all generated cases is not always useful. In order to enable better control of the retaining process, various approaches for selecting cases to be retained have been developed (Reinartz & Roth, 2000). These approaches often imply a reorganization of the entire

case base when adding a new case, for example, by removing other cases. In order to develop the prototype of CBR system the researcher use four major CBR tasks (retrieval, reuse, revise and retain).

Case-based Reasoning system advantage and disadvantage

CBR system is appropriate in medicine for some important reasons; in a similar way to physicians make reasoning about the patients and also use their expertise and show synergistic knowledge. According to (Prentzas, J. & Hatzilygeroudis, I, 2007; Pal, S., and Shiu, K., 2004) the major advantages of CBR system from different points of view, reducing the knowledge acquisition task, avoiding repeating mistakes made in the past, Providing flexibility in knowledge modeling, Making predictions about the probable success of a given solution, learning over time, reasoning in a domain with a small body of knowledge, Reasoning with incomplete or imprecise data and concepts. Providing a means of explanation, extending too many different purposes and reflecting human reasoning.

The disadvantage of CBR in the medical field, regarding the fact that a large number of features (symptoms) can be found in medical records, this makes case adoption problematic. Although the reliability of a CBR system increased with a range of problems that covers, it is not guaranteed (Hamid T et al, 2015). Feature extraction- desire to let medical CBR systems handle increasingly complex data formats, such as image, sensor signals etc. And also a limited number of available cases- in the initial phase of a medical CBR system there are often limited number of cases available. This may reduce the performance of the system. If past cases are missing or very sparse in some areas the accuracy is reduced (Shahina, B et al, 2009). However, Case-based reasoning (CBR) is now considered as a suitable technique for diagnosis and treatment in the medical domain (Rainer & Lothar, 2000).

2.5. Case-Based System Evaluation Method

Evaluation is a broad concept. Its objective is to assess a knowledge-based system overall value. In addition to assessing acceptable performance levels, it analyzes whether the system would be usable, efficient, and cost-effective. Evaluation of the knowledge-based system using a test case needs experts as evaluators. The knowledge-based system testing procedure carried out by system evaluators to classify the test cases into correct or incorrect classes. The evaluation was

done by comparing the system test result with the physician answers (as the human expert did). Therefore, System evaluators and knowledge engineers made decisions by comparing the system test result with the physician's answers. The result of the comparison shows that our approach has made a close decision as the human expert did (Siraj, 2019).

Evaluation of the knowledge-based system includes both system performance (statistical analysis) and user acceptance (Buchanan and Forsythe; 1991). The statistical analysis for case-based reasoning can be conducted for both retrieval and reuse process. The first task of case-based reasoning is to retrieve cases that are relevant to the new case (Agnar, A & Enric, P, 1994). As the retrieval task of the case-based reasoning system aims to retrieve relevant cases from the case base, precision and recall are useful measures of retrieval performance in case-based reasoning. Recall is defined as the ratio of the number of relevant cases returned to the total number of relevant cases for the new case in the case base. Whereas precision is the ratio of the number of relevant cases returned to the total number of cases for a given new case (McSherry, 2001).

To assure the applicability of the system in real life in addition to system performance evaluation based on statistical analysis the system is evaluated with user acceptance testing. Because of a system that achieves better system performance statistically may not be comfortable to the user in solving a particular problem (Teshome.M, 2015). There are varieties of methods to assess user acceptance of a knowledge-based system. Some of the most commonly used methods include interviews, checklist questions, log studies, reaction studies and visual interaction. Among these checklist questions are the most common method which allows the experts or domain users to make comments while interacting with the system and hence in this study also the researcher used it for user acceptance system evaluation

2.6. Case-Based System Development Tools

A Case-Based Reasoning (CBR) tool is software that can be used to develop several applications that require case-based reasoning methodology. There are different types of tools that can be used for developing a CBR system. Accordingly, (Essam, A & AbdEl-Badeeh, S, 2008; Iqbal & Ashraf , 2006) identified the following CBR tools.

- **ReCall:** -This case-based reasoning tool is written in C++ language. It provides both the Nearest Neighbor and inductive retrieval algorithm. It can run on Windows and UNIX workstations under Motif, Sun, HP series 700 and DEC Alpha, designed in open architecture that allows the user to add case-based reasoning functionality in the applications.
- **ReMind:** - produced by Cognitive Systems Inc., was developed with support from the US DARPA program. It was originally developed for the Macintosh and has since been ported to MS Windows and various UNIX platforms. ReMind offer template, nearest neighbor, inductive and knowledge-based retrieval. Its limitation is retrieving speed. Nearest neighbor is very slow, on the other hand, inductive retrieval is very fast. When it creates an inductive index, then it becomes slow (Watson & Marir, 1994).It will able to access data in ODBC-compliant databases and very influential tool.
- **CasePower:** -Inductive Solutions Inc. developed the CasePower tool. That tool builds its cases in a matrix environment provided by Microsoft Excel. Rows and columns of a spreadsheet are used to define cases and their attributes. It uses nearest neighbor retrieval and reduces the search time by calculating the index in advance. If the new case is retained, then entire set of case indices must be recalculated (Watson & Marir, 1994)
- **CASPAIN:** - This is a CBR tool written in C language which can run on operating systems like;MS-DOS, MAC, or UNIX with no graphical user interface. It performs simple nearest-neighbor matching to retrieve cases from the database. Store cases including adaptation rules, in the form of an ASCII file.
- **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 automatic generation of user interface. This framework is developed by the GAIA artificial intelligence group at Completeness University in Madrid (Shadia, 2018). The framework is built in two hierarchical levels-upper and lower. The lower level consists of a library of classes (Software modules) for full 4REs CBR cycle, also for the definition of cases, attributes and connectors for access to outer databases. The upper level is a “black box” graphical interface, which allows non-complicated user CBR application generation based on lower-level modules.

JCOLIBRI supports full CBR cycle. The design of the JCOLIBRI framework comprises a hierarchy of Java classes and a number of XML files. It support Nearest Neighbor retrieval algorithm. JCOLIBRI is aimed at CBR system designers. A CBR application can be built by instantiating the framework, or through the GUI-based configuration tools, which allow one to build the application without writing a line of code. And also designed as a wide spectrum framework able to support several types of CBR systems from the simple nearest-neighbor methods based on flat or simple structures to more complex Knowledge Intensive ones (Belen D et al, 2007). There are lots of CBR applications, developed on JCOLIBRI based: additional shells (abstract levels) for distributed CBR systems, statistical CBR systems, multi-agent supervisor systems for text file classification, and lots of CBR recommender systems. As a result of this jCOLIBRI is used to develop the case base reasoning system for the study.

2.7. Data Mining

Data mining is the process of extracting hidden knowledge from data and it can reveal the patterns and relationships among large amount of data in a single or several datasets (Mu-Jung, H et al, 2007). In other words data mining is one of the steps of knowledge discovery for extracting implicit patterns from vast, incomplete and noisy data. Knowledge discovery from databases is defined as the process of identifying valid, novel, potentially useful and ultimately understandable patterns of data. One of the crucial steps in Knowledge discovery is Data Mining and often they are used as synonyms (Deshpande, M.P & Thakare, D, 2010).

In medical domain data mining works on the bases of data that has been already collected and find the best possible solution by analyzing and identifying the frequent pattern or trends of past data. In medical science past experience plays a vital role in diagnosing any new situation. Basically the aim of using Data mining technique in medical domain is to facilitate hospitals, clinics, physicians, and patients by adopting new technologies, which will help in early detection of life threatening diseases, reducing treatment costs and increasing the survivability of the patient (Surabhi & Seema, 2014).

2.7.1. Data Mining Process Models

There are different DM process model standards. KDD process (Knowledge Discovery in Databases) and CRISP-DM (Cross Industry Standard Process for Data Mining) are some of the models that are used in different DM projects.

2.7.1.1. The KDD Process Model

The basic task of KDD is to extract knowledge (or information) from lower level data (databases). There are several formal definitions of KDD, all agree that the intent is to harvest information by recognizing patterns in raw data. Let us examine definition proposed by Fayyad, Piatetsky Shapiro and Smyth, "Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The goal is to distinguish from unprocessed data, something that may not be obvious but is valuable or enlightening in its discovery. Extraction of knowledge from raw data is accomplished by applying Data Mining methods (Nwagu, C et al, 2017). Generally, KDD has five phases. These are selection, preprocessing, transforming, data mining and interpretation (Ana & Manuel, 2008; Gonzalo and Oscar, 2010). The researcher used Knowledge Discovery in Database (KDD) process model to automatically acquire knowledge from the JUSH pneumonia dataset using Waikato Environment for Knowledge Analysis (WEKA) data mining tool.

Selection - this stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed; Pre-processing - this stage consists on the target data cleaning and preprocessing in order to obtain consistent data. Transformation-this stage consists on the transformation of the data using dimensionality reduction or transformation methods; Data Mining - this stage consists on the searching for patterns of interest in a particular representational form depending on the DM objectives (usually prediction); Interpretation/Evaluation - this stage consists on the interpretation and evaluation of the mined patterns (Ana & Manuel, 2008).

2.7.1.2. CRISP-Data Mining Process Model

The cross-industry standard process for data mining (CRISP-DM) is also another well-known process model to develop Data Mining projects and was proposed by a consortium of companies include of Teradata, SPSS (ISL), Daimler-Chrysler and OHRA. CRISP-DM defines the processes and tasks that you have to do in order to develop a successful Data Mining project. As it was

mentioned before, since its introduction in 1996, CRISP-DM has been the most favored methodology in data mining domain. CRISP-DM is widely applicable in industry areas. CRISP-DM also defines for each phase the tasks and the deliverables for each task. (Ahmad, N et al, 2011)

- ☛ **Business understanding:** This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
- ☛ **Data understanding:** The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, 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 data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool) from the initial raw data. 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:** this is the fourth phase of CRISP-DM process. In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.
- ☛ **Evaluation:** At this stage built a model that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results.
- ☛ **Deployment:** Model construction is generally not the end of the project. Even if 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 customer Classifier can use it.

2.7.2. Data mining Classification Techniques and Algorithms

In data mining, classification is one of the most vital task. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes (Shelly and Anand, 2011). Classification is the derivation of a function or model which determines the class of an object based on its attributes. A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. Such a classification function or model can be used to classify future objects and develop a better understanding of the classes of the objects in the database (Sivanandam and Sumathi, 2006). Classification has numerous applications including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis. Data classification is two-step process, consisting of learning step (where a classification model is constructed) and classification step (where the model is used to predict the class label for the given data) (Yaswanth & Korrapati, 2016).

Generally, Classification is a supervised data mining method applied to datasets containing an expert labeling in the form of a categorical attribute, called a class. Classification is a process of building model that define data class and used to forecast the class of objects whose class label is unknown. It finds out the connection between predictor value and the target value. The model is based on the analysis of a set of training data. The data historical, for a classification is typically divided into two datasets: one for building the model; the other for testing the model. Thus the various classification approaches can be employed on pneumonia case for obtaining specific information. Common classification techniques are neural networks, K-nearest neighbor, the naïve Bayes technique, decision trees, support vector machines and rule based learning (Hossein & Behrouz, 2016; Surabhi & Seema, 2014). In this report, decision tree, Byes classifier and rule based learning are discussed.

2.7.2.1. Decision Tree

In data mining, a decision tree (it may be also called Classification Tree) is a predictive model which can be used to represent the classification model. The use of decision trees is very popular in data mining due to its simplicity and transparency. Decision trees are tree-shaped structures that represent decision sets. These decisions generate rules, which then are used to classify data (Omkar, 2014). This structure mainly contains a starting node (called root) and group of branches (conditions) that lead to other nodes until we reach leaf node that contain final decision of this route.

According to wendwesen (2016), there are two main types of decision trees that are based on the target variable. These are classification trees and regression trees. Classification trees are decision trees used to predict categorical/discrete variables that are divided into categories. For example, the categories can be yes or no. And, the second one is regression trees, which is a decision tree used to predict continues variables. Classification trees can provide the confidence to correctly classify the data. In this case, the classification tree reports the class probability, which is the confidence that a record is in a given class. On the other hand, regression trees estimate the value of a target variable that takes on numeric value.

A decision tree is a flowchart-like tree structure that has three types of nodes (Radhwan,G et al, 2017).

- ✚ Root node: it has no incoming link and zero or more outgoing edges.
- ✚ Internal nodes: it has one incoming edge and two or more outgoing links.
- ✚ Leaf/terminal nodes: it has no outgoing link sand exactly one link incoming.

J48 classification algorithm

J48 is the WEKA (Waikato Environment for Knowledge Analysis) implementation of C4.5, which, to date, is still one of the most used in Data mining algorithms when it comes to classification. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. J48 classifier build a decision tree for the given data set, whose nodes represent discrimination rules acting on selective features by recursive partitioning of data using depth- first strategy (Mohammad et al, 2013).

The algorithm used each attribute of the data to make decision by splitting the data into smaller subjects. All the possible tests are considered during decision making based on information gain value of each attribute (Himani & Sunil, 2016).

2.7.2.2. Rule Based Classification

Though the decision tree is a widely used technique for classification purposes, another popular alternative to decision trees is classification rules which can be expressed as paths IF-THEN rules so that humans can understand them easily and represent information or knowledge in a very simple and effective way (Teseahun, 2012). A rule-based classifier uses a set of IF-THEN rules for classification; it is a relationship between antecedent, and consequent i.e. an expression of the form IF condition THEN the conclusion. The rule-based inference engine is constructed on the concept that IF the information supplied by the user satisfies the conditions of a rule, THEN the actions of the rule are executed.

Rule based classification are constructed in two ways; direct method and indirect method. Direct or sequential methods are those that extract rules directly from data, for example RIPPER. Indirect methods are those that extract rules from other classification model like decision trees e.g. C4.5 rules (Thangaraj & Vijayalakshmi, 2013). Direct methods first grow a single rule (Rule growing) then remove instances from this rule (Instance Elimination) after that prune the rule (Stopping Criterion and Rule Pruning) and then finally add rules to current rule set. PART and JRIP are algorithms which are rule based classifiers.

PART

Part is a separate-and-conquer rule learner. The algorithm producing sets of rules called decision lists which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in every iterative and makes the “best” leaf into a rule. The algorithm is a combination of C4.5 and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) rule learning (Vaishali et al, 2014; Abdi, 2016).

JRip

JRIP is a propositional rule learner. JRip proposed a Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is an inference and rules--based learner (RIPPER) that can be used

to classify elements with propositional rules. The RIPPER algorithm is a direct method used to extract the rules directly from the data. JRip (Weka's implementation of the RIPPER rule learner) is a fast algorithm for learning "IFTHEN" rules. Like decision trees rule learning algorithms are popular because the knowledge representation is very easy to interpret (Abdi, 2016)

2.7.2.3. Bayesian Network Classifiers

Bayesian classifier is statistical classifier and a practical learning algorithm that can predict class membership probabilities. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes and classification is based on a probabilistic model specification; i.e. it can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classifiers are graphical models which are very useful for representing variables (as nodes of the graph) and the probabilistic relationships between them (as connections, or edges of the graph). Bayesian networks can have different advantages. Among those, some of them provide probabilistic output, can work with limited sensor data availability, more flexible relative to engineering development than traditional expert systems, used for both data qualification (state recognition) and anomaly reasoning, can work in a central or distributed run-time environment either shore-side or shipboard. The reason why use Bayesian networks is Bayesian inference methods have proven to be valuable for knowledge-based data mining applications, and are based on a causal (explanation based) modeling framework. Because relationships between variables in a Bayesian network are defined probabilistically, trends can be detected and analyzed over a continuous scale, rather than in a Boolean fashion. These classifiers are used in many fields and one common class of classifiers are Naive Bayes classifiers (Jiawei, 2006)

Naive Bayes

Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions which assumes all of the features are equally independent. One of the most effective Bayesian classifiers, in the sense that its predictive performance is competitive with other classifiers, is the Naive Bayesian classifier. This classifier learns from training data the conditional probability of each attribute A_i given the class label C . Classification is then done

by applying Bayes rule to compute the probability of C given the particular instance of A_1, \dots, A_n , and then predicting the class with the highest posterior probability (Wendwesen, 2016)

For this study Naïve Bayes, PART, and J48 classification algorithms are used because of their effectiveness and efficiency in order to build the predictive model. According to Daniel (2013), J48 decision tree algorithms can be applied on discrete, continues and categorical data and to get a simple rule, which allow the study on different data types to get a better result. Anbarasi et al. (2010), states Naive Bayes algorithm has a good accuracy and speed to build the predictive model on large data as compared as other classification algorithms. PART classifier algorithm support all type of classes like binary and nominal class and supports all type of attributes (Vaishali et al, 2014). After performing experimentations on those classification algorithms the best one has been selected for building the model for the case based system.

2.8. Related Works

In the domain of health, different researchers in different Universities and research centers throughout the world have been conducted medical diagnosis and treatment knowledge-based systems in the past decades.

For similar problem Abebayehu (2015) was developed a user friendly CBR system for diagnosis and treatment of bacterial pneumonia and viral pneumonia diseases. This main motivation of his study was lack sufficient experts and most of cases are treatable if there is supportive system. To do so, knowledge engineering research design method were employed to achieve the overall objective. The necessary data was acquired from Bahir Dar Felege Hiwot Referral Hospital using interview and document analysis. To model and represent the acquired knowledge decision tree and rule based reasoning was employed. As a result, the developed system using ProLog and Java perform 83.33% and 90.33 system performance and user acceptance respectively. However, the proposed study aim to developed CBR system using data mining techniques to enhance the system performance and user acceptance. Furthermore, the current study amid to include other types of pneumonia disease for under five year children which is more critical and prevalent currently.

For diagnosis and treatment of tuberculosis (TB) chronic infection disease, Amelework (2017) develop a case based reasoning system due the lack of sufficient medical experts in Ethiopia. To

do so, the appropriate knowledge was acquired using both structured and unstructured interview with medical experts and literature reviews. The acquired knowledge was modeled through CommonKADS and represented using CBR techniques. The developed system was scored 74% of recall and 83% precision. The developed system performance was also measure through user acceptance testing which scored 86%. However, the proposed study aims to used data mining techniques in order filter out best medical cases before implementing CBR system, which is mainly enhance performance of the CBR system. Since, Amelework (2017) system performance can be enhanced using data mining techniques to acquire best cases in CBR system database.

Zhenjia, Liangping and Runfeng(2020) was conduct a research for comparison and validation of different deep learning model in order to diagnosis of Pneumonia. The main motivation to conduct the study was the necessity of diagnosis and treatment of Pneumonia at early stage and which time consuming task for medical experts. To do so, they collect a 5216 train and 624 test chest X-ray images having normal and pneumonia classes available data from Kaggle online data source and models are implemented using Python. As a result from various experimentation conducted with different machine learning algorithms, deep neural network scored better accuracy by improving Mobile Net's network structure. However, the current work focused on acquiring already solved cases to design a CBR system as knowledge sharing tool during diagnosis and treatment of Pneumonia.

Another interesting was conducted to localized knowledge based system for diagnosis and treatment of pediatric Pneumonia was also conducted by Melquiades and Haile (2019). Their study was amid to solve the shortage of skilled medical experts in the area and the problem of language for diagnosis of treatment of Pneumonia patients. To solve that, the researchers develop a localized KBS using rule based reasoning method using (SWI) Prolog tool. The proper knowledge was acquired from St.Marry Hospital Axum, Axum University Referral Hospital and Axum Health Centers. Melquiades and Haile (2019) used decision tree to model the knowledge and production rule were used to represent it. As a result, the localized KBS perform 87.5% overall accuracy and 88% users are satisfied with the localized pediatric pneumonia knowledge-based system. Whereas the proposed study including known domain knowledge with actually solved cases which improved the overall performance of system and user acceptance of the system.

Aiyesha et al(2019) also conducted a study for differentiate the diagnosis of Tuberculosis and Pneumonia using machine learning algorithms. to improve the final outcome of the study , the proper data preprocessing tasks such as handling missing value, useless value and discretization continues values was undertaken. A total of 705 cases having 32 attributes with two classes are used for experimentation. In order determine the diagnosis for those two diseases, they used Gaussian Naïve Bayes, decision tree and random forest classification algorithms in the experiment. As a result of their study, Gaussian Naïve Bayes score 92.9%, decision tree score 93.85% and random forest scored 97.64% of overall accuracy. However, these study was conducted for choosing the best classification model in order to differentiate Tuberculosis and Pneumonia diseases which differ with the current prosed study. To fill the applicable system after selecting the best classification model the, the researcher aimed to develop CBR prototype system for ease of use by end users.

Another interesting work also conducted by Ermiyas and Hailemicheal (2020) using CBR techniques for diagnosis of chronic kidney diseases. The researchers was motivated to conducted this investigation due lack qualified medical experts and sufficient medical equipment for diagnosis and treatment of chronic kidney diseases. Therefore, the knowledge was acquired though interviewing medical experts from Jimma university referral hospital, St. Paulose millennium hospital and Hawassa university referral hospital. While, the appropriate case features are extracted/generated for representation. The main reason CBR was the ability of reuse of the new cases of chronic kidney diseases for future unlike rule based system. The researchers used JCOLIBRI case based framework tool for developing the prototype system. The proposed system used similar system development tool, whereas the researcher collect solved cases for the documented classes form Jimma University specialized hospital and taking best cases using data mining classification algorithms in order enhance the system performance.

In order to use the benefits of best retrieval stages of cases, Lucky ,Endang and Much(2017) used one of CBR case retrieval method called nearest neighbor algorithm in order to develop an expert system for diagnosis of bowel disease. To implement the expert system the actual 60 recorded medical cases are collected and used. As a result, the system perform 95% accuracy, which better as compared with other pervious. However, when number of cases used for implementing are reduced indirectly highly increased the system performance. In this context the

researcher aims at using large amount of dataset and filter out best cases using data mining techniques rather than directly using un-processed datasets. Since, by increasing number of collected cases diagnosis different causality it's possible to increases the system acceptance by end users. Moreover the researcher used JCOLIBRI to implement the system for effective use of new knowledge or cases in the future.

In order to use the advantages of data mining techniques for prediction or classification various causality of medical diseases, Bezahegn (2017) used J48 and PART data mining techniques to develop a predictive model for pre-diabetics screening. The necessary 4529 data for mining purpose was collected from Adere general hospital in Hawassa city Ethiopia and used Cross-Industry Standard Process of Data Mining (CRISP-DM) process model for archived the overall objective of the study. As result, PART prediction model scored 96.78 prediction accuracy. And also User acceptance testing performs an average 92% acceptance was achieved. however his has some limitation on adapting new knowledge or reusing new cases, to solve this the issues the current proposed work used CBR techniques to reuse of new cases. Moreover, the proposed study used Naïve Bayes classification algorithm in order to solve model building time performance.

Recently, a diet calorie determination system using case based reasoning conducted by Hindayati et al. (2020) for the unbalance food consumption problem. The cases used for system implementation was collected from online available data sets which includes age, gender, height and activity attributes. The researcher was also adapted retrieve, reuse, revise and retain techniques for developing a CBR system in order to solve the problem. One of this study contribution was it developed the CBR system having user interface which better for system usability by end user. The result of the study also show retrieving similar cases form cases based on their increasing order. However, including all cases without filtering consume data storage capacity. To fill this gap, the proposed study used different data mining techniques in order to select best cases and reduced number repeated cases.

Kedir (2018) uses data mining classification technique, especially decision tree for the diagnosis and treatment of diabetes disease by applying experimental research design. The general objective of this study is to design and develop prototype knowledge based system using data mining techniques for diagnosis and treatment of diabetes. They use three data mining

classification algorithm these are J48, PART and JRip. Finally, decided to use the results of J48 classification algorithm score 95.1515%, because it registered better performance than other classifiers. They used Weka for model construction and evaluation, Ultimate Visual basic studio 2013(Vb.net) for using data mining results as store knowledge base and as front side of prototype and common lisp prolog (Clisp) used for obtained knowledge backend coding. The system performance is 92% and user acceptance testing perform 91.43% acceptance was achieved. As their recommendation to be able the system refine the knowledge base. And also not generate the classification model through command on Weka “SimpleCLI” which ensure each taking cases accuracy to classify diabetes disease. As a result, the current study is different from this research by “SimpleCLI” which ensure each taking cases accuracy to classify disease.

Mekdes (2018) conducted a research by using a Case-based reasoning approach for diagnosis malnutrition only under five-year children. The general objective of this study was to design a case-based reasoning system that provides expert advice for the diagnosis of malnutrition under five-year children. The knowledge was acquired from Jimma University specialized hospital and Hawasa university comprehensive specialized hospitals and design science were followed to design a prototype case-based reasoning system. Stratified sampling technique was employed to select domain experts for knowledge acquisition and for system testing and evaluation from Jimma University specialized hospital. The researcher used a hierarchical tree manner of knowledge modeling. Evaluation of a knowledge base system includes both system performance (statistical analysis) and user acceptance. The statistical analysis for CBR can be conducted for both retrieval and reuse process. Based on evaluating the performance of the system, the average precision and recall values achieved were 71% and 83% respectively. And also User acceptance testing performs an average 83% acceptance was achieved. For the development of the prototype system, the researcher used jCOLIBRI version 1.1 implementation tools and nearest neighbor algorithm used.

Another study on application of data mining techniques conducted by Desalegn (2017) to develop knowledge based system to determine the choice of contraceptive methods using data mining technique. The general objective of this study is to develop knowledge based system to determine choice of contraceptive methods. The researcher used hybrid data mining model for the data mining task, Rule based knowledge representation approach to represent knowledge,

Prototyping approach to develop the knowledge based system, WEKA to mine hidden knowledge, Microsoft visual basic.Net programming language to codify the represented rules in knowledge base. To build a predictive model it uses the classification algorithms namely J48 decision tree, JRIP rule, REPTree, PART and Naïve Bayes. On the classification J48 is selected where scored 72.3% accuracy, because it registered better performance than other classifiers by percentage split (66% / 34%). Based on system performance evaluation and user acceptance test, 86.6 % of accuracy and 76% acceptance was scored.

Moreover, the following table shows summarized related works with their achievement

<i>Author and year</i>	<i>Title of the work</i>	<i>Used tools</i>	<i>Result achieved</i>	<i>Significance</i>
Abebayehu (2015)	A user friendly knowledge-based system for diagnosis and treatment of pneumonia	RBR, ProLog and Java	90.33 User acceptance	Used as for diagnosis and treatment of pneumonia
Zhenjia, et al (2020)	Comparison and Validation of Deep Learning Models for the Diagnosis of Pneumonia	Python	Not specified	To support experts for diagnosis of pneumonia
Melquiades and Haile (2019)	localized knowledge based system for diagnosis and treatment of pediatric pneumonia: the case of tigray central zone in Ethiopia	RBR, ProLog	88% users are satisfied	To solve the shortage of skilled medical experts in the area and the problem of language for diagnosis of treatment of Pneumonia patients
Aiyasha et al (2019)	Differential Diagnosis of Tuberculosis and Pneumonia	Not specified	Not	To support experts for differentiate the diagnosis of

	using Machine Learning		specified	Tuberculosis and Pneumonia
Amelework (2017)	Application of case based reasoning for tuberculosis diagnosis	CBR, jCOLIBRI	86%. user acceptance	To support experts for diagnosis of tuberculosis
Ermiyas and Hailemicheal (2020)	Chronic Kidney Disease Diagnosis Model Based on Case Based Reasoning	CBR, jCOLIBRI	Not specified	To support experts for diagnosis of Chronic Kidney Disease
Lucky et al. (2017)	Expert System Diagnosis of Bowel Disease Using Case Based Reasoning with Nearest Neighbor Algorithm	CBR, Not specified	95% system accuracy	To support experts for diagnosis of diagnosis of Bowel Disease
Bezahegn (2017)	Developing a predictive model for pre- diabetes screening by using data mining technology	WEKA	PART 96.78% prediction accuracy	To support experts for pre- diabetes screening
Hindayati et al. (2020)	Diet Calorie Determination System using Case-Based Reasoning	CBR	Not specified	It used to determining a calorie diet per day for each person with similarity values based on case-based.
Kedir (2018)	Developing knowledge based system using data mining techniques for diagnosis and treatment of diabetes	RBR, WEKA, Vb.net, Prolog	J48 score 95.1515%, 91.43% user acceptance	To support experts for diagnosis and treatment of diabetes
Mekdes (2018)	A case based reasoning system for diagnosis of	CBR,	83% User	To support experts for diagnosis

	malnutrition for under-five year children	jCOLIBRI	acceptance score	malnutrition under-five year children
Desalegn (2017)	Developing knowledge based system to determine the choice of contraceptive methods using data mining technique	RBR WEKA, Vb.net	J48 score 72.3%. 76% user acceptance	To determine choice of contraceptive methods

Different researches have been conducted to apply a knowledge-based system in a supportive medical domain. Moreover, Amelework (2017), Ermiyas & Hailemicheal (2020), Lucky et al. (2017), Hindayati et al. (2020) and Mekedes (2018) have been developed using case-based representation technique to reason out the solution of a particular problem. And also Ababayehu (2015) and Melquiades and Haile (2019) have been developed using rule-based representation technique

But, the developed KBS were not used automatic knowledge extraction from datasets using data mining techniques. Based on the related works reviewed the researcher attempts to apply a knowledge-based system by using data mining techniques for diagnosis and treatment of pneumonia on under-five year children. For this reason, the researcher attempted to design case-based reasoning system by using data mining techniques to identify suitable cases for diagnosis and treatment of pneumonia for under-five year children.

CHAPTER THREE

METHODOLOGY

3. Methodology of the Study

Research methodology is a set of systematic technique used in research. This simply means a guide to research and how it is conducted. It describes and analysis methods, throws more light on their limitations and resources, clarify their pre-suppositions and consequences, relating their potentialities to the twilight zone at the frontiers of knowledge (Chinelo, 2016).The methodology refers to the research design, reviewing literature, procedures, tool and techniques followed. This research will be designed to develop a prototype CBR system by using a data mining technique that provides expert advice for diagnosis and treatment of pneumonia under five-year children. Here under, the following methods and technique are discussed in detail to achieve the aim of the study.

Research Design

In this study, design science research approach is used to design a case-based system for diagnosis and treatment of pneumonia under-five year children. For this study, predictive data mining task mainly classification techniques and KDD model is applied for generating representative cases from the prepared data. The required dataset for data mining purpose were acquired from Jimma University specialized hospital. In order to make knowledge extraction as much as correct as possible (i.e. in order to keep the correctness of the knowledge as it is kept at the source) different techniques could be applied. Among these techniques, data mining techniques and, more general, KDD techniques became the most used in the recent years. KDD is the process of extracting and refining useful knowledge from large data (Mihaela K. , 2006).

Design science research aims to improve the understanding of information systems phenomena by creating information technology artifacts. The artifacts created embody the solution for a problem previously defined (Myers, Lawrence, & Tuunanen, 2017). According to Peffers et al (2008) design science is an outcome-based information science research methodology, which offers special guidelines for evaluation and iteration within the research design. Design sciences research is both, a process of developing new solutions to existing problems and matching

existing solutions to new problems (Weber, 2012). The process is organized in three main phases “problem identification”, “solution design” and “evaluation” that can relate with each other within the research process (Philipp, O et al, 2009). For this reason design science is the appropriate research design for this study which aimed primarily at discovering artifacts and solving problem as opposed to accumulation of theoretical knowledge. Due to these reasons, the researcher selected design science research approach for this study.

Figure 3.1 depicts the phases of design science research and discussed as follows:

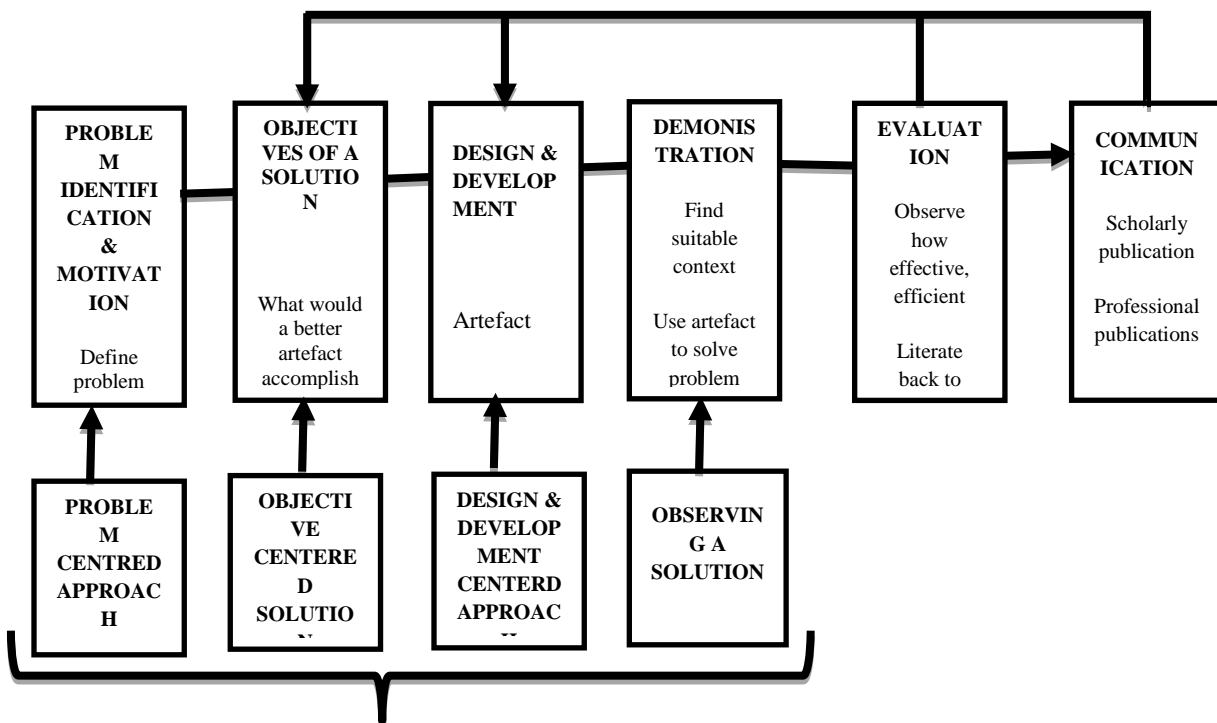


Figure 3 1:- Design Science research process model adopted from (Peffer, K et al, 2008)

3.1. Problem identification and Motivation

As the clearly stated in the statement of the problem section, the researcher has done different techniques to identify some existing problems in the area in order to achieve the objectives of the study. The researcher read different journals articles, books and manuals, conference papers, reports and other scholarly communicated materials. After reading those materials, the researcher observed that pneumonia under five year children is one of the top killing diseases according to world health organization report of (2019) which estimates 18% of all deaths of children under five years old worldwide due to pneumonia disease.

Study Area

The study area of this research was Jimma University Specialized Hospital which exists in Oromia regional state of Ethiopia, in the city of Jimma. The researcher has select this organization with the assumption of one of the oldest public hospitals in Ethiopia and it gives medical service for all patients who come from the southwest region of Ethiopia. So it has experienced experts and it is an educational hospital in Ethiopia. Therefore, lots of recorded documents in order to get the required information on pneumonia under five year children was possible.

Sampling Techniques

In this study, a purposive sampling technique was used to select domain experts for knowledge acquisition from Jimma University specialized hospital based on their level of experiences and availability. It is one of the most common sampling techniques in qualitative research and it's appropriate to capture demonstrable experience and expertise of the experts. The criterion to select domain experts for the study was by considering their professions, educational qualification and years of experience in the diagnosis and treatment of pneumonia under-five year children. Initially, to get some overview of pneumonia for under five-year children different interviews were conducted with eight experts that include pediatrician, doctors and nurses to consult and get suggestions about the diseases as well as the treatments. Unstructured interview questions were prepared and forwarded to the selected domain experts for acquiring important knowledge.

The researcher has been used both manual and automatic knowledge acquisition mechanisms. Knowledge Discovery Databases (KDD) model has been used to acquire knowledge from the Jimma University specialized hospital dataset (cases) using WEKA data mining tool. Knowledge Discovery in Databases denotes the task of revealing significant relationships and regularities in data base on the use of algorithms collectively entitled "data mining". The researcher use available cases with the help of experts for the research to get a better result and to increase the efficiency and effectiveness of the prototype. In addition the researcher also reviewed different domain related articles, pneumonia diagnosis manuals, document analysis that has been used by reading different books, journals articles, different previous researches work, websites, recorded

cards which are related to pneumonia disease to better understand and to support the domain knowledge. This enabled the researcher to understand the dominant attributes and the structure of the data.

3.2. Objectives of the solution

After collecting the required knowledge from previously solved pneumonia cases, domain experts and document analysis, the next step is setting objectives of the solution. The main objectives of the solution is to show a way how to diagnosis and treat pneumonia under five year children by using data mining as a knowledge acquisition technique and using it's result for designing a case based reasoning system which would enable the medical practitioners, pediatrician, nurses and other health professionals to consult and get suggestions about the diseases as well as the treatments.

3.3. Design and Development Approaches for Knowledge Use

The design and development of prototype system involve knowledge acquisition and representation. In this study predictive data mining task and KDD model is applied for generating representative cases from the prepared data. Specially, classification technique is used to build representative cases used to design the case based reasoning prototype for diagnosis and treatment pneumonia under five year children.

First, the collected data was prepared and best describing attributes were selected in accordance to the objective of the study. In this stage is all about determining the attributes and missing values to be filled, smoothing noises, recognizing outliers and correcting inconsistent. Data format is also transformed to ARFF file format which is the suitable file format for the data mining tool. Secondly, the researcher conducted three experiments for the classification by employing Data Mining algorithms such as J48 pruned, PART and naïve Bayes.

Finally, the researcher compared the result of these classification algorithms performance after experimentation. PART rule classifier algorithm conducted through 80/20% percentage splitter model training and testing option had 98.44% accuracy which is better than the remaining two algorithms. Therefore, based on the objective classifier algorithm evaluation criteria's, the researcher decide to use PART rule classifier algorithm. Likewise, classification models that are developed in this research are evaluated using a test dataset based on their classification accuracy

and interpretation also made accordingly. As a result a test instance which registers more than 99 % accuracy was taken as a knowledge source for CBR development.

In the process of knowledge-based system development, knowledge modeling is one of the basic steps. Knowledge modeling involves organizing and structuring of the knowledge gathered during knowledge acquisition. Knowledge modeling is the concept of representing information and the logic for purpose of capturing, sharing and processing knowledge to simulate intelligence. Here, the basic concepts that reveal the main activities and decisions that are made to solve cases in the domain are modeled (Henok, 2011). There are different techniques that can be used in modeling the domain knowledge, for example, decision tree and hierarchical tree structure.

Decision tree scan help steps (decisions) to find a solution for a certain problem domain and identify various ways of splitting a data set into branch like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. Decision trees are an important tool for decision making, and risk analysis which are usually represented in the form of a graph or list of rules. And also support models that classify patterns using a sequence of well-defined rules (Jelena,D et al, 2013). The hierarchical tree diagram provides the analyst with an effective visual condensation of the clustering results. The hierarchical tree diagram is one of commonly used methods of determining the number of clusters. It is also useful in spotting outliers, as these will appear as one member clusters that are joined later in the clustering process (Hemant & Limaye, 2011). For this study, the researcher used a decision tree knowledge modeling technique. Decision tree structure can easily model concepts and clearly explains the concepts in the problem area at hand.

Knowledge representation is one of the basic steps in the process of knowledge-based system development. There are many different methods of knowledge representation: semantic net, rules, frames and cases are the most popular method of knowledge representation currently (Solomon, 2013). For this study, the case-based knowledge representation method has been used because it clearly demonstrates the domain knowledge. Case-based reasoning is a process that uses similar problems previously mitigated to solve the current problem. The necessary cases and the knowledge from the domain expert and different relevant documents were acquired and modeled. The next task is coding the knowledge into computer using appropriate and efficient

knowledge representation methods. Therefore, the researcher employed a case based representation method for this research project.

The major objective in this phase is to take the acquired knowledge and translate it into machine-readable form using various knowledge representation techniques. For this research the researcher represented the knowledge from the manual knowledge acquisition mechanisms through conceptual modeling, data mining results as rules and used feature-value case representation for case based development. This approach uses old experiences to understand and solve new problems. It also reuses its solutions and lessons learned for future use. In addition, it represents cases in an easy way by using attribute and value pair representation. The algorithms used to calculate the similarity of cases in a case base representation for this research was nearest neighbor retrieval algorithm. The similarity function of nearest neighbor retrieval algorithm involves in computing the similarity between the stored cases in the case base and the new query. After that, it selects the most similar stored cases to the query (Tamir, A et al, 2017)

3.4. Demonstration

In this study system demonstration is used to show the efficiency of the artifact to solve the problem and how to use the artifact to solve the problem. It also enables the targeted group to have a concise understanding on how the designed system functions and to give feedbacks for the researcher. According to (Tamir, A et al, 2017; Antanassov, A & Antonov, L., 2012), there are various KBS development tools which are available both freely and commercially, Among this SWI-prolog, myCBR, and jCOLIBRI are among the most widely used and known frameworks for teaching and academic research purpose. All of the above-mentioned tools have their own capabilities and limitations, jCOLIBRI framework has the following features. A CBR tool could be used to develop several applications that require case based reasoning methodology. The major advantages of jCOLIBRI as compared to other implementation tools includes the following such as jCOLIBRI supports the full CBR cycles such as (Retrieval, Reuse, Revise and Retain) (Iqbal & Ashraf , 2006; Tamir, A et al, 2017). Hence in this study for the development of CBR prototype system, the researcher used JCOLIBERI version 1.1 which is object oriented framework.

3.5. Evaluation Methods

After developing CBR prototype, it was tested its functionality and user acceptance of the system. The evaluation processes focus on system's user acceptance of the prototype and the performance of the system. The researcher used Precision, Recall, F-measure and True Positive rate to evaluate the results and accuracy of the data mining model. The researcher also evaluated the KBS using system performance testing by preparing test cases and users' acceptance testing of the system by using visual interaction methods together with questionnaire which helps the researcher to make sure that whether the potential users would like to use the proposed system frequently and whether the proposed systems meets user requirements.

CHAPTER FOUR

KNOWLEDGE ACQUISITION, MODELING AND EXPERIMENTATION

4. Knowledge Acquisition

Knowledge acquisition (KA) is an essential part of developing a KBS using the suitable methods that should be used for acquiring relevant knowledge from domain experts and other sources of information such as patient card, books, databases, guidelines, manuals, journal articles and computer files. The development of an efficient knowledge-based system (KBS) involves the development of an efficient knowledge base that has to be complete, clear and non-redundant, but at the same time it is the most difficult one that needs great care, patience and attention in the stage of case base development.

According to Asghar & Iqbal (2009) knowledge extraction should be as much as correct as possible (i.e. in order to keep the correctness of the knowledge as it is kept at the source) different techniques could be applied. Among these techniques, data mining techniques and more general knowledge discovery techniques became the most used in the recent years. In this research, the researcher acquired the knowledge using two types of knowledge acquisition methods which are manual and automatic knowledge acquisition using data mining techniques base line medical datasets were collected from JUSH.

4.1. Manual Knowledge Acquisition

Pneumonia is an infection of the lung tissue. It can affect one or both lungs. The lung tissue is made up of thin-walled sacs that contain air. When a person has pneumonia their air sacs become filled with microorganisms, fluid and inflammatory cells and their lungs are not able to work properly. Pneumonia is the single leading cause of mortality in children under five and is a major cause of child mortality in every region of the world, with most deaths occurring in sub Saharan Africa and South Asia. Pneumonia kills more children under five than AIDS, malaria, and measles combined, yet increased attention in recent years have been on the latter diseases. Pneumonia is a form of acute respiratory tract infection (ARTI) that affects the lungs. When an individual has pneumonia, the alveoli in the lungs are filled with pus and fluid, which makes breathing painful and limits oxygen intake (WHO, 2020). According to Harrison's textbook of

internal medicine defines pneumonia as an infection of the pulmonary parenchyma caused by various organisms. It states that pneumonia is not a single disease but a group of specific infections, each with a different epidemiology, pathogenesis, presentation and clinical course (Grant, 2016)

According to American Lung Association (2020 May) symptoms of pneumonia vary from mild to severe, depending on factors such as the type of germ causing the infection, and your age and overall health. Common clinical symptoms of pneumonia can range from mild to severe include (JohnsHopkins, 2020; Cedars-Sinai, 2020)

- Cough, which may produce greenish, yellow or even bloody mucus,
- Fever, Excessive sweating and shaking chills
- shortness and fast of breathing
- Sharp or stabbing chest pain that gets worse when you breathe deeply or cough
- Loss of appetite, low energy, and fatigue
- Nausea and vomiting, especially in small children
- Headache, Fussiness and Wheezing
- Muscle pain, Blue skin, Skin Rash and Dizziness
- Difficulty swallowing and convulsion

In order to prevent pneumonia in children is an essential component of a strategy to reduce child mortality. Immunization against Hib (Homophiles influenza type b), pneumococcus, measles and whooping cough (pertussis) is the most effective way to prevent pneumonia. Adequate nutrition is key to improving children's natural defenses, starting with exclusive breastfeeding for the first 6 months of age, to more specific infection control measures like hand-washing, avoiding individuals with signs of respiratory tract infections, and vaccinations. In addition to being effective in preventing pneumonia, it also helps to reduce the length of the illness if a child does become ill. And also addressing environmental factors such as indoor air pollution (by providing affordable clean indoor stoves, for example) and encouraging good hygiene in crowded homes also reduces the number of children who fall ill with pneumonia (WHO, 2020; UNICEF, 2019).

4.1.1. Treatment and Types of Pneumonia

Types of pneumonia are referred to by the type of organism that causes the inflammation, such as bacterial pneumonia, viral pneumonia, or fungal pneumonia and other organism (WHO, 2020; JohnsHopkins, 2020)

☛ Bacterial pneumonia

This type of pneumonia is caused by various bacteria. It usually occurs when the body is weakened in some way, such as by illness, poor nutrition, old age, or impaired immunity, and the bacteria are able to work their way into the lungs. Bacteria can go down into your lungs. When this happens, the air sacs in your lungs get infected and inflamed. They fill up with fluid, and that causes pneumonia. Bacterial pneumonia can affect all ages, but you are at greater risk if you abuse alcohol, smoke cigarettes, are debilitated, have recently had surgery, have a respiratory disease or viral infection, or have a weakened immune system (JohnsHopkins, 2020).

Symptoms may be a bit different for each child. They may also depend on what is causing the pneumonia. Cases of bacterial pneumonia tend to happen suddenly with these symptoms: Cough that produces mucus, Cough pain, Vomiting or diarrhea, Chest pain, Loss of appetite, Tiredness (fatigue), Fever, Fast or hard breathing, Wheezing (Cedars-Sinai, 2020; WHO, 2020)

Table 4.1:- Bacterial pneumonia treatment

Place of acquired	Common symptom	Age	Treatment
HAP	<ul style="list-style-type: none">- Cough- Chest pain- Vomiting or diarrhea- Loss of appetite- Tiredness	2 months up to 12 months	<ul style="list-style-type: none">➤ Ampicillin: 350 mg IM/IV every six hours for at least five days (At Hospital)➤ Gentamicin: 80mg IM/IV once a day for at least five days (At Hospital)
		12 months up to 3 years	<ul style="list-style-type: none">➤ Ampicillin: 600 mg IM/IV every six hours for at least five days (At Hospital)➤ Gentamicin: 100mg IM/IV once a day for at least five days (At Hospital)

	<ul style="list-style-type: none"> - HGF - Fast or hard breathing - Wheezing - Convulsion 	3 years up to 5 years	<ul style="list-style-type: none"> ➤ Ampicillin: 850 mg IM/IV every six hours for at least five days (At Hospital) ➤ Gentamicin: 130mg IM/IV once a day for at least five days (At Hospital)
CAP	<ul style="list-style-type: none"> ➤ Chest in drawing or Fast breathing. ➤ Fever ➤ Cough 	2 months up to 12 months	➤ Amoxicillin dispersible tablets (250 mg) 1 tab twice a day x 5 days (10tabs) (At Health center and Health post)
		12 months up to 3 years	➤ Amoxicillin dispersible tablets (250 mg) 2 tabs twice a day x 5 days (20 tabs) (At Health center and Health post)
		3 years up to 5 years	➤ Amoxicillin dispersible tablets (250 mg) 3 tabs twice a day x 5 days (30 tabs) (At Health center and Health post)

☛ Viral pneumonia

Viral pneumonia is an infection of your lungs caused by a virus. The most common cause is the flu, but you can also get viral pneumonia from the common cold and other viruses. These nasty germs usually stick to the upper part of your respiratory system. But the trouble starts when they get down into your lungs. Then the air sacs in your lungs get infected and inflamed, and they fill up with fluid. Viral Pneumonia Spreads through the air in droplets of fluid after someone sneezes or coughs. These fluids can get into your body through your nose or mouth. You can also get viral pneumonia after touching a virus-covered doorknob or keyboard and then touching your mouth or nose (Carol, 2018)

Viral pneumonia usually moves in steadily over a few days. On the first day it feels like the flu, with symptoms like: Cough that produces mucus, Vomiting or diarrhea, Loss of appetite, Tiredness (fatigue), LGF, Chills, Fast or hard breathing, Headache and Fussiness (Cedars-Sinai, 2020; Carol, 2018).

Table 4.2:- Viral pneumonia treatment

Place of acquired	Common symptom	Age	Treatment
HAP	<ul style="list-style-type: none"> ➤ Cough that produces mucus ➤ Cough pain ➤ Vomiting or diarrhea ➤ Loss of appetite ➤ Tiredness (fatigue) ➤ LGF ➤ Chills ➤ Fast or hard breathing ➤ Headache ➤ Fussiness 	2 months up to 12 months	<ul style="list-style-type: none"> ➤ Oseltamivir 25mg twice daily for 5 days (At Hospital) ➤ Home care advice:- Soothe the throat and relieve the cough with a safe remedy
		12 months up to 3 years	<ul style="list-style-type: none"> ➤ Oseltamivir 35mg twice daily for 5 days (At Hospital) ➤ Home care advice:- Soothe the throat and relieve the cough with a safe remedy
		3 years up to 5 years	<ul style="list-style-type: none"> ➤ Oseltamivir 50mg twice daily for 5 days (At Hospital) ➤ Home care advice:- Soothe the throat and relieve the cough with a safe remedy
CAP	<ul style="list-style-type: none"> ➤ Cough that produces mucus ➤ Cough pain ➤ Vomiting or diarrhea ➤ Loss of appetite ➤ LGF ➤ Chills ➤ Fast or hard breathing ➤ Headache 	2 months up to 12 months	<ul style="list-style-type: none"> ➤ Oseltamivir 25mg twice daily for 5 days days (At Hospital) ➤ Home care advice:- Soothe the throat and relieve the cough with a safe remedy
		12 months up to 3 years	<ul style="list-style-type: none"> ➤ Oseltamivir 35mg twice daily for 5 days days (At Hospital) ➤ Home care advice:- Soothe the throat and relieve the cough with a safe remedy

		3 years up to 5 years	<ul style="list-style-type: none"> ➤ Oseltamivir 50mg twice daily for 5 days (At Hospital) ➤ Home care advice:- Soothe the throat and relieve the cough with a safe remedy
--	--	-----------------------	--

☛ **Aspiration pneumonia**

Aspiration pneumonia is an inflammation of your child's lungs. It may have happened after your child breathed in (aspirated) a foreign substance. This could be a substance such as food, liquid, vomit or saliva are inhaled and cause inflammation in the lungs. Aspiration may have happened because your child has a health problem that makes it hard to swallow normally. Aspiration can often be prevented by dietary interventions for dysphagia e.g. adjusting texture, consistency and amount of food and fluids, frequent oral care and post-pyloric tube (tube passes through stomach and into small intestine) feedings. Aspiration can cause signs and symptoms in a baby such as chest pain, HGF, cough, possibly with green sputum, blood, or a foul, Shortness of breath, wheezing, fatigue (Tiredness), blue skin, difficulty swallowing, Faster breathing while feeding and Excessive sweating (Cedars-Sinai, 2020)

Table 4.3:- Aspiration pneumonia treatment

Place of acquired	Common symptom	Age	Treatment
HAP	<ul style="list-style-type: none"> ➤ Chest discomfort ➤ High Grade fever. ➤ Cough, possibly with green sputum, blood, or a foul ➤ Shortness of breath ➤ Wheezing ➤ Fatigue (Tiredness) ➤ Blue skin 	2 months up to 12 months	<ul style="list-style-type: none"> ➤ Put on Mechanical ventilation (At Hospital) ➤ <u>Clindamycin</u> 220 mg IV every 8 hours (followed by 110 mg orally 4 times/day) and <u>amoxicillin/clavulanate</u> 230 mg IV every 12 hours for 1 to 2 weeks. (At Hospital)

	<ul style="list-style-type: none"> ➤ Difficulty swallowing ➤ Faster breathing while feeding ➤ Excessive sweating 	12 months up to 3 years	<ul style="list-style-type: none"> ➤ Put on Mechanical ventilation (At Hospital) ➤ <u>Clindamycin</u> 360 mg IV every 8 hours (followed by 180 mg orally 4 times/day) and <u>amoxicillin/clavulanate</u> 360 mg IV every 12 hours for 1 to 2 weeks. (At Hospital)
		3 years up to 5 years	<ul style="list-style-type: none"> ➤ Put on Mechanical ventilation (At Hospital) ➤ <u>Clindamycin</u> 510 mg IV every 8 hours (followed by 255 mg orally 4 times/day) and <u>amoxicillin/clavulanate</u> 510 mg IV every 12 hours for 1 to 2 weeks. (At Hospital)
CAP	<ul style="list-style-type: none"> ➤ Chest discomfort ➤ High Grade fever. ➤ Cough, possibly with green sputum, blood, or a foul ➤ Shortness of breath ➤ Wheezing ➤ Fatigue (Tiredness) ➤ Blue skin ➤ Difficulty swallowing ➤ Faster breathing while feeding ➤ Excessive sweating 	2 months up to 12 months	<u>Clindamycin</u> 220 mg IV every 8 hours (followed by 110 mg orally 4 times/day) and <u>amoxicillin/clavulanate</u> 230 mg IV every 12 hours for 1 to 2 weeks. (At Hospital)
		12 months up to 3 years	<u>Clindamycin</u> 360 mg IV every 8 hours (followed by 180 mg orally 4 times/day) and <u>amoxicillin/clavulanate</u> 360 mg IV every 12 hours for 1 to 2 weeks. (At Hospital)

		3 years up to 5 years	<u>Clindamycin</u> 510 mg IV every 8 hours (followed by 255 mg orally 4 times/day) and <u>amoxicillin/clavulanate</u> 510 mg IV every 12 hours for 1 to 2 weeks. (At Hospital)
--	--	-----------------------	--

☛ Mycoplasma pneumonia

Mycoplasma pneumonia (MP) is a contagious respiratory infection that spreads easily through contact with respiratory fluids. It can cause epidemics. MP is known as an atypical pneumonia and is sometimes called “walking pneumonia.” It spreads quickly in crowded areas, such as schools, college campuses, and nursing homes. When an infected person coughs or sneezes, moisture containing the MP bacteria is released into the air. Uninfected people in their environment can easily breathe the bacteria. Patients present with symptoms of upper respiratory tract infection, cough, LG fever, tiredness, skin Rash, chest or stomach pain, vomiting and wheezing (Graham, 2018).

Table 4.4:- Mycoplasma pneumonia treatment

Place of acquired	Common symptom	Age	Treatment
HAP	<ul style="list-style-type: none"> ➤ Cough ➤ cough that may produce some mucus ➤ LG fever ➤ Tiredness ➤ Headaches ➤ Skin Rash ➤ chest or stomach pain ➤ vomiting ➤ wheezing 	2 months up to 12 months	<ul style="list-style-type: none"> ➤ Prednisolone 5mg once daily for 5 day ➤ Erythromycin 80mg IV QID for 5 day (At Hospital)
		12 months up to 3 years	<ul style="list-style-type: none"> ➤ Prednisolone 5mg once daily for 5 day ➤ Erythromycin 125mg IV QID for 5 day (At Hospital)

		3 years up to 5 years	<ul style="list-style-type: none"> ➤ Prednisolone 5mg once daily for 5 day ➤ Erythromycin 170mg IV QID for 5 day (At Hospital)
CAP	<ul style="list-style-type: none"> ➤ Cough ➤ cough that may produce some mucus ➤ LG fever ➤ Tiredness ➤ Headaches ➤ Skin Rash ➤ chest or stomach pain ➤ vomiting ➤ wheezing 	2 months up to 12 months	<ul style="list-style-type: none"> ➤ Prednisolone 5mg once daily for 5 day ➤ Erythromycin 80mg PO QID for 5 day (At Health center)
		12 months up to 3 years	<ul style="list-style-type: none"> ➤ Prednisolone 5mg once daily for 5 day ➤ Erythromycin 125mg PO QID for 5 day (At Health center)
		3 years up to 5 years	<ul style="list-style-type: none"> ➤ Prednisolone 5mg once daily for 5 day ➤ Erythromycin 170mg PO QID for 5 day (At Health center)

☛ Fungal pneumonia

Fungal pneumonia is most common in people with chronic health problems or weakened immune systems, and in people who are exposed to large doses of certain fungi from contaminated soil or bird droppings. Symptoms of fungal pneumonia is (fever, cough, headache, rash, muscle aches, or joint pain) are similar to other common illnesses, diagnosis and treatment are often delayed. In a very small proportion of people, the infection can cause chronic pneumonia, spread from the lungs to the rest of the body and cause meningitis (brain or spine infection), or even death (CDC, 2012).

Table 4.5:- Fungal pneumonia treatment

Place of acquired	Common symptom	Age	Treatment
HAP	<ul style="list-style-type: none"> ❖ Fever ❖ Cough ❖ Headache ❖ Skin rash ❖ Muscle pains 	2 months up to 12 months	➤ Fluconazole 60mg PO every 72 hours (At Health center)
		12 months up to 3 years	➤ Fluconazole 100mg PO every 72 hours (At Health center)
		3 years up to 5 years	➤ Fluconazole 150mg PO every 72 hours (At Health center)
CAP	<ul style="list-style-type: none"> ❖ Fever ❖ Cough ❖ Headache ❖ Skin rash ❖ Muscle pains 	2 months up to 12 months	➤ Fluconazole 60mg PO every 72 hours (At Health center)
		12 months up to 3 years	➤ Fluconazole 100mg PO every 72 hours (At Health center)
		3 years up to 5 years	➤ Fluconazole 150mg PO every 72 hours (At Health center)

☛ **Broncho pneumonia**

Bronchopneumonia is a type of pneumonia, a condition that causes inflammation of the lungs. The bronchi are the large air passages that connect the windpipe to the lungs. These bronchi then split into many tiny air tubes known as bronchioles, which make up the lungs. At the end of the bronchioles are tiny air sacs called alveoli where the exchange of oxygen from the lungs and carbon dioxide from the bloodstream takes place. Pneumonia causes an inflammation in the lungs that leads to these alveoli filling with fluid. This fluid impairs normal lung function, producing a range of respiratory problems. Bronchopneumonia is a form of pneumonia that affects both the alveoli in the lungs and the bronchi (Aaron, 2018). Symptoms of broncho pneumonia may include fever, shortness of breath, chest pain that may get worse with coughing or breathing deeply, coughing up mucus, sweating, chills or shivering, muscle aches, tiredness, loss of appetite, headaches, dizziness, nausea and vomiting (Graham, 2018; Aaron, 2018)

Table 4.6:- Broncho pneumonia treatment

Place of acquired	Common symptom	Age	Treatment
HAP	<ul style="list-style-type: none"> ❖ Fever ❖ shortness of breath ❖ Chest pain ❖ Coughing up mucus ❖ Sweating ❖ Chills or shivering ❖ Muscle aches ❖ Tiredness ❖ Loss of appetite ❖ Headaches ❖ Dizziness ❖ Nausea ❖ Vomiting ❖ Coughing up blood 	2 months up to 12 months	<ul style="list-style-type: none"> ➤ Ampicillin: 350 mg IM/IV every six hours for at least five days (At Hospital) ➤ Gentamicin: 80mg IM/IV once a day for at least five days (At Hospital)
		12 months up to 3 years	<ul style="list-style-type: none"> ➤ Ampicillin: 600 mg IM/IV every six hours for at least five days (At Hospital) ➤ Gentamicin: 100mg IM/IV once a day for at least five days (At Hospital)
		3 years up to 5 years	<ul style="list-style-type: none"> ➤ Ampicillin: 850 mg IM/IV every six hours for at least five days (At Hospital) ➤ Gentamicin: 130mg IM/IV once a day for at least five days (At Hospital)
CAP	<ul style="list-style-type: none"> ❖ Fever ❖ shortness of breath ❖ Chest pain ❖ Coughing up mucus ❖ Sweating ❖ Chills or shivering ❖ Muscle aches ❖ Tiredness ❖ Loss of appetite ❖ Headaches ❖ Dizziness ❖ Nausea ❖ Vomiting ❖ Coughing up blood 	2 months up to 12 months	<ul style="list-style-type: none"> ➤ Ceftriaxone 350mg IV once daily for 7 day (At Hospital)
		12 months up to 3 years	<ul style="list-style-type: none"> ➤ Ceftriaxone 600mg IV once daily for 7 day (At Hospital)
		3 years up to 5 years	<ul style="list-style-type: none"> ➤ Ceftriaxone 850mg IV once daily for 7 day (At Hospital)

☛ **Ventilator-associated pneumonia (VAP)**

Ventilator-associated pneumonia (VAP) is a lung infection that develops in a person who is on a ventilator. A ventilator is a machine that is used to help a patient breathe by giving oxygen through a tube placed in a patient’s mouth or nose, or through a hole in the front of the neck. An infection may occur if germs enter through the tube and get into the patient’s lungs. CDC provides guidelines and tools to the healthcare community to help end ventilator-associated pneumonia and resources to help the public understand these infections and take measures to safeguard their own health when possible. Symptoms of VAP may include fever, chills, cough, and shortness of breath, chest pain and coughing up mucus. VAP prevention process measures are now better established and many are supported by randomized controlled trials. Preventive strategies are aimed at avoiding unnecessary intubation, decreasing the duration of ventilation, preventing aspiration, and minimizing inoculation and colonization of the lower respiratory tract with mouth, gastrointestinal and upper respiratory tract flora (Morrow,B et al, 2008; Andrew, R et al, 2009)

Table 4 7:- Ventilator-associated pneumonia treatment

Place of acquired	Common symptom	Age	Treatment
HAP	❖ Fever ❖ chills ❖ Cough ❖ Shortness of breath ❖ Chest pain ❖ Coughing up mucus	2 months up to 12 months	➤ Ceftriaxone 350mg IV once daily for 7 day (At Hospital)
		12 months up to 3 years	➤ Ceftriaxone 600mg IV once daily for 7 day (At Hospital)
		3 years up to 5 years	➤ Ceftriaxone 850mg IV once daily for 7 day (At Hospital)
CAP	❖ Fever ❖ chills ❖ Cough ❖ Shortness of breath ❖ Chest pain ❖ Coughing up mucus	2 months up to 12 months	➤ Ceftriaxone 350mg IV once daily for 7 day (At Hospital)
		12 months up to 3 years	➤ Ceftriaxone 600mg IV once daily for 7 day (At Hospital)
		3 years up to 5 years	➤ Ceftriaxone 850mg IV once daily for 7 day (At Hospital)

☛ Streptococcus pneumoniae

Streptococcus pneumoniae (pneumococcus) is a Gram-positive bacterium that is responsible for the majority of community-acquired pneumonia. It is a commensal organism in the human respiratory tract, meaning that it benefits from the human body, without harming it. However, infection by pneumococcus may be dangerous, causing not only pneumonia, but also bronchitis, otitis media, septicemia, and meningitis. People with pneumococcal disease can spread the bacteria to others when they cough or sneeze. Symptoms of pneumococcal infection depend on the part of the body affected. Symptoms can include fever, cough, shortness of breath, chills and fatigue (Fleck, 2019)

Table 4.8:- Streptococcus pneumonia treatment

Place of Acquired	Common symptom	Age	Treatment
HAP	❖ Fever ❖ Chills ❖ Cough ❖ Shortness of breath ❖ Fatigue	2 months up to 12 months	➤ Ceftriaxone 350mg IV BID for 7 day (At Hospital)
		12 months up to 3 years	➤ Ceftriaxone 600mg IV BID for 7 day (At Hospital)
		3 years up to 5 years	➤ Ceftriaxone 850mg IV BID for 7 day (At Hospital)
CAP	❖ Fever ❖ Chills ❖ Cough ❖ Shortness of breath ❖ Fatigue	2 months up to 12 months	➤ Ceftriaxone 350mg IV BID for 7 day (At Hospital)
		12 months up to 3 years	➤ Ceftriaxone 600mg IV BID for 7 day (At Hospital)
		3 years up to 5 years	➤ Ceftriaxone 850mg IV BID for 7 day (At Hospital)

4.1.3. Pneumonia Category

Pneumonia can be classified or characterized in different ways. Health care professionals often refer to pneumonia based upon the way that the infection is acquired by location, such as community-acquired pneumonia or hospital-acquired pneumonia. Based on different microbial causes and patient factors, which need different management strategies (Sinan, E et al, 2014).

Community-acquired pneumonia (CAP)

Community-acquired pneumonia (CAP): This is the most common form of pneumonia and describes pneumonia that is acquired outside of the hospital or health care environment. In most cases pneumonia is not spread from person to person and quite often is transmitted via droplets in the air, touching contaminated objects, poor hygiene and sharing cups or utensil or from the environment (CDC, 2012).

Hospital-acquired pneumonia (HAP)

Hospital acquired pneumonia (HAP) is defined as pneumonia that occurs 48 hours or more after admission in a patient who had no signs of disease at the time he or she was presenting Pneumonia in to the hospital. Acquired when an individual is already hospitalized for another condition. HAP is generally more serious because it develops in ill patients already hospitalized or under medical care for another condition

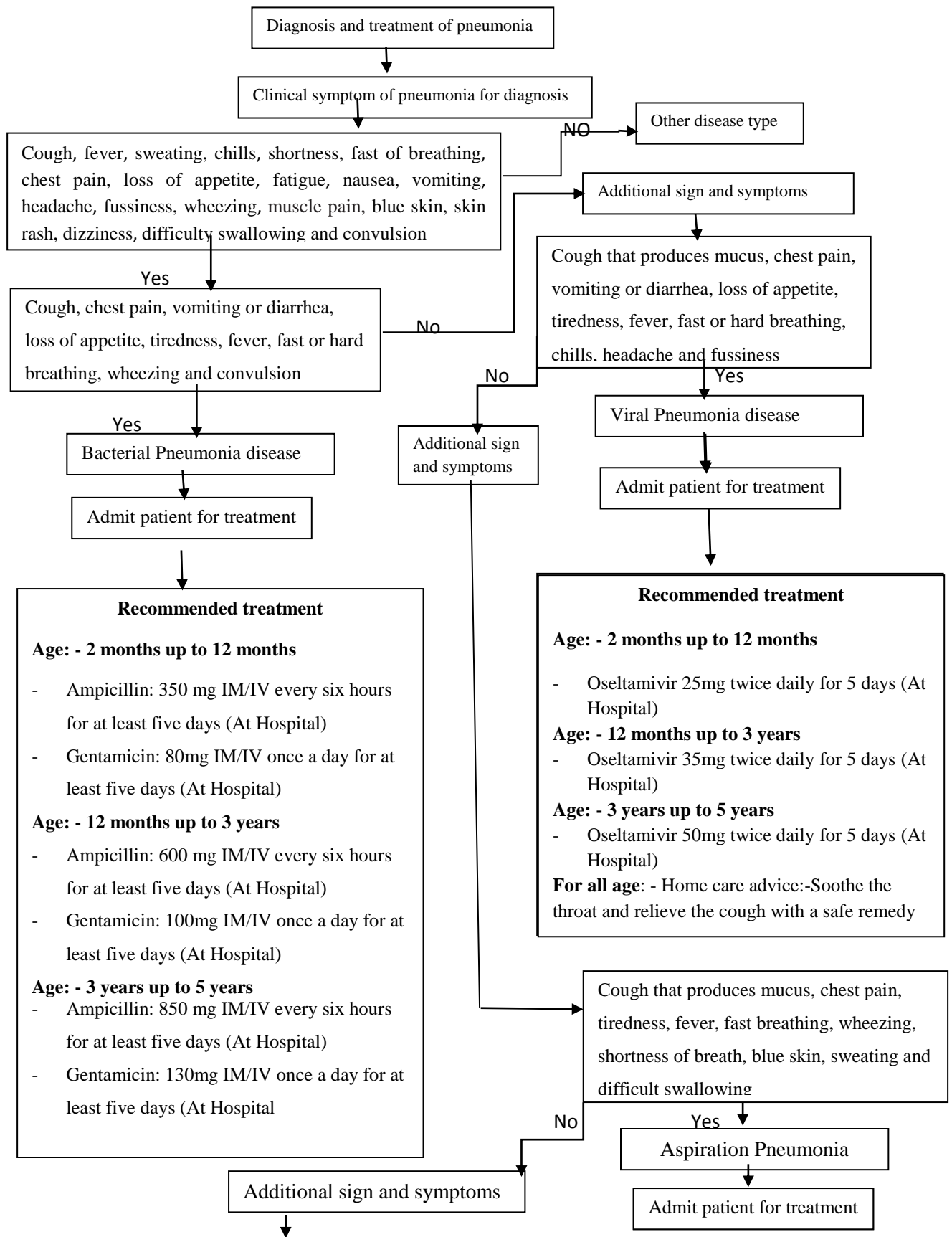
4.2. Case Modeling

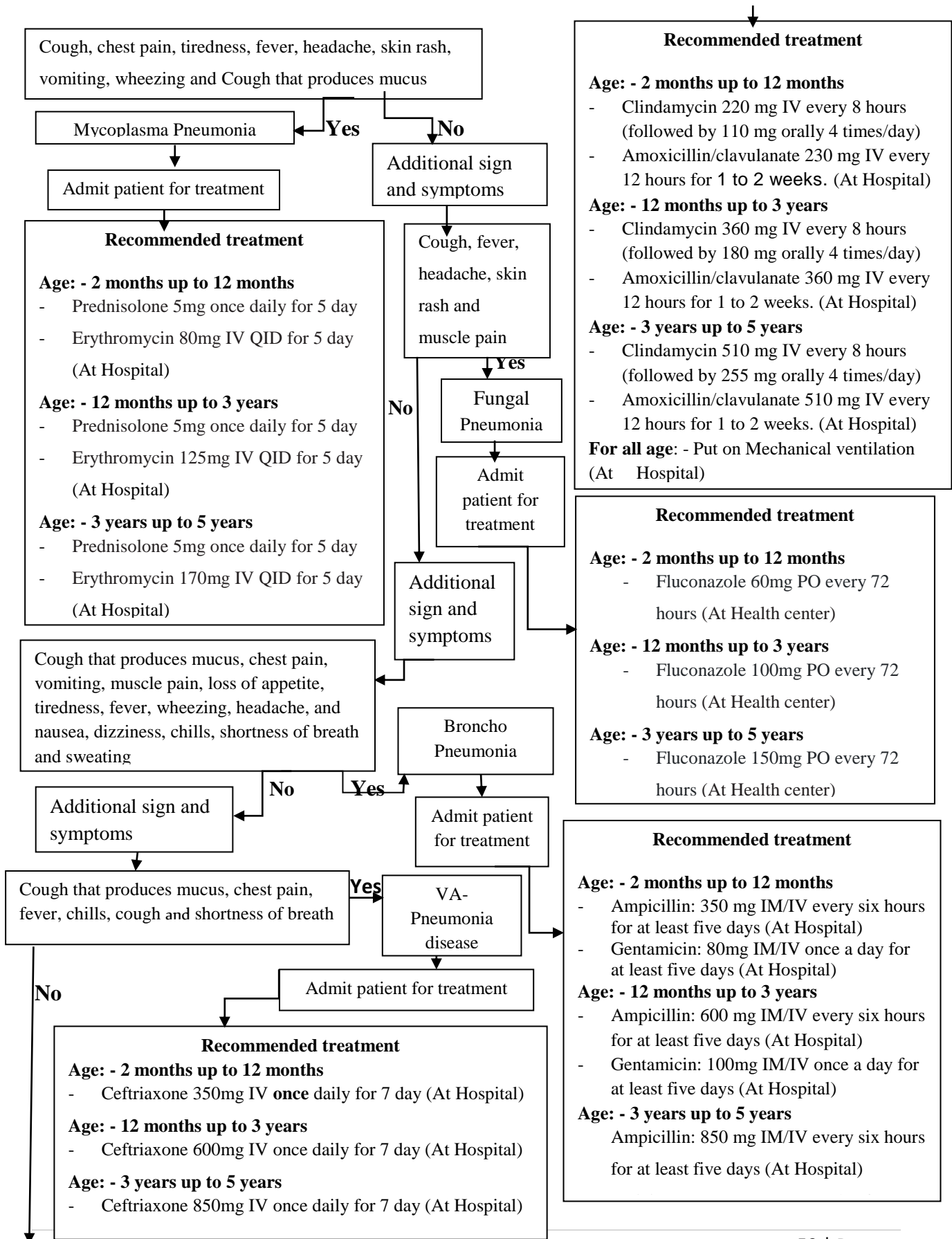
After the required case is acquired from pneumonia cases, domain experts (health professionals) and other relevant documents, the next step is modeling the case. The knowledge modeling step involves organizing and structuring of the knowledge gathered during knowledge acquisition. This activity provides an implementation independent specification of the knowledge to be represented in the knowledge base. Knowledge modeling is the concept of representing information and the logic for the purpose of capturing, sharing and processing knowledge to simulate intelligence. This model helps to ensure that all stakeholders in a proposed system understand the language and terminology being used and quickly conveys information for validation and modification where necessary (Makhfi, 2011). Here, the basic concepts that tell the main activities and decisions are made to solve cases in the domain are modeled.

Conceptual modeling is a crucial step in the knowledge acquisition process so as to understand well the problem domain and to prepare the knowledge representation phase. There are different conceptual modeling techniques and for this study decision tree structure is used to model how pneumonia diagnosis and treatment for under-five year children is performed. Because, Decision tree structure can easily model concepts and clearly explains the concepts in the problem area at hand.

4.2.1. Conceptual Modeling Using Decision Tree

In the diagnosis of pneumonia for under-five year children, the domain experts have a concept of symptoms that are used to differentiate the real symptoms of different pneumonia diseases. For those pneumonia diseases, the domain experts (health professionals) have general knowledge about the common sign and symptoms of each pneumonia disease. In the knowledge acquisition time, the domain experts (health professionals) explained that there are symptoms that are used for diagnosing the new patient who came for treatment. In addition to that in the patient cards, there are some additional identifying symptoms whether the patient has which type of pneumonia disease. The possible symptoms used for the domain experts to identify which pneumonia disease is affecting the patient are presented in the following figure 4.1. Therefore, the researcher identified the different signs and symptoms with the help of pediatrician experts.





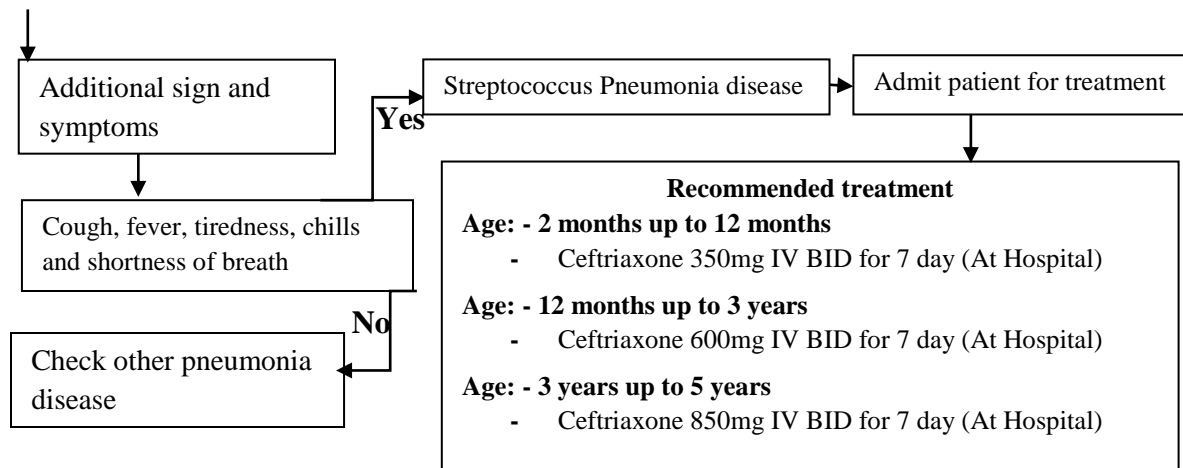


Figure 4.1:- Decision Tree for Diagnosis and Treatment of Pneumonia

4.3. Knowledge Acquired from Data Mining

Knowledge Acquisition is the process of eliciting knowledge from relevant and related sources from individual patient's card history or cases, documents, manuals and medical web sites, which helps in building complete, accurate and well organized knowledge based systems. In this study knowledge acquired using data mining by applying classification algorithms. The classifications are then used as a case for designing and developing case based reasoning system for diagnosis and treatment of pneumonia diseases.

Today's in healthcare industries generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and enhance the quality of the decision making process. Due to this tendency data mining application in healthcare sectors today is great, because healthcare organizations today are capable of generating and collecting a large amounts of data. This increase in volume of data needs automatic way for these data to be extracted when needed. With the use of data mining techniques it is possible to extract interesting and useful knowledge and these knowledge can be used by experts for efficient and enhanced decision making process (Durairaj & Ranjani, 2013; Wendwesen, 2016). Knowledge used for the designed system is extracted from the data mining tool. The data used for the data mining tool was collected from Jimma University specialized hospital. Receiving enough and necessary data is the hardest task

besides preparing the data for the data mining tool in order to achieve the intended objective of the study.

In addition, Knowledge acquisition is a complex and time-consuming stage during case based system development (Wendwesen, 2016). For case generation and model building, classifier algorithms such as Naïve Bayes, PART, and J48 are employed and their result is compared to generate best rules and representative model for the case based system.

After collecting the data, generating significant and representative data from the data set is a crucial task. In this stage, the researcher generates 1614 instances with corresponding attributes are collected for this study from Jimma university specialized hospital. In this manner some data set information are unnecessary from the instance list. In the preprocessing, Weka tool is used to replace missing values. Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data. After processing the data, 1614 instances with 25(including the class attribute) selected attributes were used. For this study, the researcher used data from the latest year 2019 back to 2016. The recent data were used due to the dynamic nature of the disease, diagnosis and treatment advancement over years. As time are changed, diagnosis technologies are advanced and drugs dosage are changed, due to this reason nearly the recent time data are used in this study.

4.3.1. Data preprocessing

Today's real-world data are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results (Soumen, C et al, 2009). Therefore, prior to giving the data to a data mining tool, preprocessing of the data is necessary. Data mining tools need well prepared data to perform the targeted tasks designed by the researcher. In data mining, the data preparation is responsible for identifying quality data from the data provided by data pre-processing systems. Preprocessing the data includes multiple steps to assure the highest possible data quality, thus efforts are made to detect and remove errors, resolve data redundancies, and taking into account of the patient privacy, to remove patient identifiers.

There are different data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store. Data transformations, such as normalization, may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features. These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to common format. Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining (Jiawei, 2006). In this study the researcher performs preprocessing activities to make the data more suitable for data mining techniques.

4.3.1.1. Data cleaning

Raw data may have incomplete records, noisy values, outliers and inconsistent data. Data cleaning(or data cleansing) routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies in the data (Deshpande, M.P & Thakare, D, 2010; Jiawei, 2006). Anticipating that data will be 100% complete and error free is unrealistic when working with patient data which collected in complex health care systems. Cleaning the data is proved a nontrivial and tedious task. Data error identification is both an automated and a manual process, and required an iterative procedure that drew upon expertise from the clinical experts as well as statistical experts and the data warehouse engineer (Inderpal, 2013).

The researcher were cleaned the data that has been collected from Jimma university specialized hospital according to the pneumonia under five-year children patient’s baseline data by discussing domain experts. In line with this some attributes which are believed unnecessary for the decision making process by the experts are removed prior to the data preprocessing task. Those attributes with their respective reasons are mentioned below in Table 4.9.

Table 4.9:- Removed attributes

Removed attributes		
No	Attributes name	Reason
1	Patients name	Not necessary

2	Patients card number	Not necessary
3	Address	Not necessary

4.3.1.2. Attribute selection

In this study, the researcher removed some attributes that has no contribution to the diagnosis and treatment of pneumonia under five-year children that is mainly related to the organization and used for only organizational data management purpose. Attribute selection is defined as “the process of finding a best subset of features, from the original set of attributes in a given data set, which is optimal according to the defined goal and criterion of attribute selection (attribute goodness criterion) (Getachew, B et al, 2017). In this case the researcher closely worked with domain experts to identify and reject attributes of those related diseases. Based on this attributes like Patient name, card number and address are the removed attributes.

After removing unnecessary attributes 25(including the class attribute) selected attributes were used in this study. As table 4.10 depicts the selected attributes with their description.

Table 4.10:-Selected attributes with their description

N_o	Attributes	Description of attributes	Attributes values
1	Sex	The sex of the patient	F-Female M-Male
2	Age	The age of the patient	Numeric values
3	Cough that produces mucus	A productive cough is when you have a cough that produces mucus or phlegm sputum	True or False
4	Cough	clinical symptoms	True or False
5	Vomit	dislodging the food in stomach through mouth	True or False
6	Diarrhea	increase in the frequency of bowel movements or a decrease in the form of stool	True or False
7	Loss of appetite	a decreased appetite occurs	True or

		when somebody has a reduced desire to eat	False
8	Tiredness	It is the condition where a person feels of reduced or no energy,	True or False
9	Fever	It holds patients body temperature	NGF, LGF or HGF
10	Chills	The patient feel cold	True or False
11	Fast breathing	the process of fast moving of air into and out of the lungs	True or False
12	chest pain	Any discomfort around the chest	True or False
13	Headache	It is one of the symptoms is shown in the pneumonia patients. This attribute used to determine whether the patient has a headache or not	True or False
14	Wheezing	The shrill whistle or coarse rattle you hear when your airway is partially blocked	True or False
15	Difficulty swallowing	If the patient difficulty swallowing food or liquids.	True or False
16	Nausea	It is an unpleasant, diffuse sensation of unease and discomfort, often perceived as an urge to vomit.	True or False
17	Dizziness	This attribute used to determine whether the patient has dizziness or not	True or False

18	Fussiness	A patient's mental state may be confused or delirious	True or False
19	Shortness of breathing	the patient feel that you can't catch your breath or get enough air	True or False
20	Skin rash	change of the human skin which affects its color and appearance	True or False
21	Sweating	If the patient excess sweating is due to an underlying medical condition	True or False
22	Muscle pains	the patient feel the muscle pain	True or False
23	Convulsions	The patient uncontrolled shaking of the body	True or False
24	Place	Place of acquisition pneumonia	<ul style="list-style-type: none"> - Community Acquired Pneumonia (CAP) - Hospital Acquired Pneumonia (HAP).
25	Final Classification	Status of patient	<ul style="list-style-type: none"> - Bacterial-Pneumonia - Viral-Pneumonia - Broncho-Pneumonia - Mycoplasma-Pneumonia - VA-Pneumonia - Aspiration-pneumonia - Fungal-Pneumonia - Streptococcus-pneumonia

However, the researcher performed attributes significance through information gain method after identified with domain experts. Since, the following figure 4.2 shows the output of attributes ranked from WEKA data mining tool.

Attribute selection output

```
Ranked attributes:
0.755    23 Convulsions
0.6254   13 headache
0.4113   15 Difficulty swallowing
0.3466   20 Skin rash
0.3427   10 Chills
0.3319   14 sweating
0.3175    5 Vomit
0.2958   11 Fast breathing
0.2321    6 Diarrhea
0.2317    7 Loss of appetite
0.2197   22 Muscle pains
0.1861   21 Wheezing
0.1762   24 Place
0.1621    8 Tiredness
0.1551   19 Shortnes of breathing
0.151     3 Cough that produces mucus
0.1437    2 Age
0.1429   12 chest pain
0.0954   18 Fussiness
0.0782   16 Nausea
0.0533    9 Fever
0.0504    4 Cough
0.036    17 Dizziness
0.0114    1 Sex

Selected attributes: 23,13,15,20,10,14,5,11,6,7,22,21,24,8,19,3,2,12,18,16,9,4,17,1 : 24
```

Figure 4.2:-Information Gain result for attribute selection

4.3.1.3. Data Transformation

Data transformation techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data. This leads to a concise, easy-to-use, knowledge-level representation of mining results (Inderpal, 2013; Jiawei, 2006).

Since the attribute values of “Age” and “Fever” are continuous and various, the researcher use data transformation and replace the actual data with the result to make the data more suitable for data mining. Table 4.11 showed list of attributes with their discretized and transformed value.

Table 4.11:-Discretized attributes with values

Attributes	Range	Distinct Values before data discretization	Distinct Values after data discretization
Age	2 months up to 12 months = (0,1] 12 months up to 3 years = (1,3] 3 years up to 5 years = (3,5]	28	3
Fever	36 - 37.5 ⁰ C= NGF 37.6 - 37.9 ⁰ C =LGF ≥ 38 ⁰ C = HGF	34	3

In table 4.11, patient's age and fever are discretized into three groups. This kind of discretization is made with the intention that it can make the result of the analysis more interpretable and understandable. And to make the data suitable to the technique PART selected in the Weka software. All these discretization were done by consulting the domain experts and by using range of number acceptable by the domain experts.

4.3.1.4. Data Formatting

WEKA (Waikato Environment for Knowledge Analysis) needs data to be prepared in some formats and file types. The datasets provided to this software were prepared in a format that is acceptable for Weka software. Data transformation is the mapping and conversion of data from one format to another that are suitable for the data mining tool (Manikandan, 2010). The original data is in the form of excel which is not suitable for the WEKA tool. So, it needs converting in to CSV and ARFF format. After converting the dataset into ARFF format the next step was opening the file with the Weka Data mining software. Here is below also the sample ARFF file format that the data mining algorithm is used for classifying.

Table 4.12:-Sample ARFF used for classification

```

@relation 'ALL DATA SET weka new'

@attribute Sex {M,F}
@attribute Age {'(1,3]','(0,1]','(3,5]' }
@attribute 'Cough that produces mucus' {No,Yes}
@attribute 'Cough ' {Yes,No}
@attribute Vomit {Yes,No}
@attribute Diarrhea {No,Yes}
@attribute 'Loss of appetite' {No,Yes}
@attribute 'Tiredness ' {No,Yes}
@attribute Fever {LGF,HGF,NGF}
@attribute Chills {No,Yes}
@attribute 'Fast breathing' {Yes,No}
@attribute 'chest pain' {No,Yes}
@attribute Headache {No,Yes}
@attribute ' sweating' {No,Yes}
@attribute 'Difficulty swallowing' {No,Yes}
@attribute Nausea {No,Yes}
@attribute Dizziness {No,Yes}
@attribute Fussiness {No,Yes}
@attribute 'Shortnes of breathing ' {No,Yes}
@attribute 'Skin rash' {No,Yes}
@attribute ' Wheezing' {Yes,No}
@attribute 'Muscle pains' {No,Yes}
@attribute 'Place ' {CAP,HAP}
Classification {'Bacterial-Pneumonia ','Viral-Pneumonia ','Broncho-Pneumonia,Mycoplasma-Pneumonia,VA-Pneumonia,Aspiration-pneumonia,Fungal-Pneumonia,Streptococcus-pneumoniae}

@data
M,'(1,3]',No,Yes,Yes,No,No,No,LGF,No,Yes,No,No,No,No,No,No,No,No,No,No,No,No,CAP,'Bacterial-Pneumonia '
F,'(0,1]',No,Yes,Yes,Yes,No,No,HGF,No,Yes,No,No,No,No,No,No,No,No,No,No,No,CAP,'Viral-Pneumonia '
M,'(0,1]',No,Yes,Yes,No,Yes,No,HGF,No,Yes,Yes,Yes,No,No,No,No,No,No,No,No,No,CAP,'Viral-Pneumonia '
M,'(0,1]',No,Yes,Yes,No,No,No,HGF,No,Yes,No,No,No,No,No,No,No,No,No,No,No,CAP,'Bacterial-Pneumonia '
M,'(1,3]',No,Yes,No,No,Yes,No,HGF,No,Yes,No,No,Yes,No,No,No,No,No,No,No,No,CAP,'Broncho-Pneumonia

```

4.4. Experimentation

After preparing the data in a suitable form for the data mining tool, the next task is experimenting using different algorithms. Hence, the aim of the data mining part in this study was building a predictive model for designing a case based system, using classification data mining algorithm. The sampled data set contains 1614 instances with 25 attribute (including the class attribute) and all of them are involved in all experiments. Default value of parameters in WEKA Data Mining tool for the classifier algorithms is taken into consideration, since it allows achieving better accuracy compared to modifying the default parameters values.

4.4.1 Experiment Design

Before starting experimentation setting how the prediction model is assed and evaluated is crucial. Since, in this particular study, 10-fold cross-validation and percentage splitter model training and testing option were used. For the case of 10-fold cross validation, the total dataset are fed and test the prediction model using one fold and use the rest nine folds for training, in such a way iteratively training and test the model. On the case of percentage split, the researcher used 80% of the data for model training and 20% of the data for testing the prediction model.

However, before decide split of the data for training and testing of the model, the researcher conducted various experiment by configuring default splitter setting, into 70%/30%. 80%/20% and 90%/10% options. Accordingly, the researcher decide to use 80%/20% splitter because the data collected from the original source are not that match enough and 90%/10% does not takes two classes of pneumonia on pre-experimentations.

In this study, the researcher used J48 decision tree, PART and Naïve Bayes classifier algorithms. These, three algorithm are selected based on their capability, simplicity, comparable accuracy in previous studies and robustness.

For experimentation result analysis the researcher used the following attribute name by replacing the original attribute name.

Table 4.13:-Attribute name used for experimentation analysis

Original attribute name	Replaced attribute name
Bacterial-Pneumonia	Bac-Pneu
Viral-Pneumonia	Vi- Pneu
Broncho-Pneumonia	Bro- Pneu
Mycoplasma-Pneumonia	Myc- Pneu
VA-Pneumonia	VA- Pneu
Aspiration-pneumonia	Asp- Pneu
Fungal-Pneumonia	Fun- Pneu
Streptococcus-pneumoniae	Str- Pneu

Moreover, the researcher prepared a separate test dataset using percentage split methods. Therefore, from the total 1614 data set the researcher prepared 323(20%) of the data for testing the model performance and the remaining 1291(80%) instances for training.

4.4.1.1 Experimentation One

In this particular experimentation, J48 decision tree prediction algorithm is applied through configuring 10-fold and percentage splitter model training and testing options. Since, the first experiments were conducted through 10-fold cross-validation with default parameters of WEKA data mining tool. The following table 4.14 show the confusion matrix output of the first experiment.

Table 4.14:-Confusion matrix of J48 decision tree with 10-fold cross validation

Actual class	Classified class
--------------	------------------

	Bac-Pneu	Vi-Pneu	Bro-Pneu	Myc-Pneu	VA-Pneu	Asp-Pneu	Fun-Pneu	Str-Pneu
Bac-Pneu	581	4	2	0	0	5	0	0
Vi-Pneu	2	160	0	0	0	0	1	0
Bro-Pneu	0	0	27	0	0	0	0	0
Myc-Pneu	0	0	0	82	0	0	0	0
VA-Pneu	0	0	0	0	14	0	0	0
Asp-Pneu	6	0	3	0	0	305	1	0
Fun-Pneu	0	0	0	0	0	1	50	0
Str-Pneu	0	0	0	0	1	0	0	46

The above confusion matrix output shows the experiment conducted through J48 prediction algorithm with 10-fold cross validation model training and testing option. Since, from the total 1291 instance, the algorithm classify 1265(97.98%) instances correctly, whereas the remaining 26(2.01%) of instances were incorrectly classified in to other type of pneumonia classes. Furthermore, on the above experimentation 40 number of leaves and 75 nodes (size of tree) were generated by the algorithm.

After building classifier algorithm the next step was evaluation of the model in KDD process. Therefore, the performance of the model is evaluated through some criteria's. Since, in this study the model performance is evaluated through performance accuracy, time taken to build the model, ROC curves, and true positive and false positive rate. The following table 4.15 shows the summarized performance evaluation result of J48 decision tree experimentation.

Table 4.15:-Summary of J48 decision tree classifier experiment result

Model characteristics	Experiment results
Accuracy	97.98%
Time taken	0.11 seconds
AV.TPR (%)	0.98
AV.FPR (%)	0.007
AV.PR	0.98
AV.RR	0.98
AV.ROC	0.986
F-Measure	0.98
CCI	1265
ICI	26

Key: AC: Average, TPR: True Positive Rate, FPR: False Positive Rate, PR: precision Rate, RR: Recall Rate, ROC: Relative Optical character curve, CCI: correctly classified instances, ICI: incorrectly classified instances.

As clearly shown on the above table 4.15, J48 decision tree algorithm classified 1265(97.98%) instances correctly and 26(2.01%) incorrectly from the given 1291 instances. J48 also takes 0.11 seconds in order to build the model and register 98.4% TP rate. Moreover, J48 decision tree algorithm had 98.6% performance on ROC curve performance measure.

In order to properly compare and select the best classifier model, the researcher proposed two model training and testing options. Since, once finished classifier model through default parameters, the next task is conducting of another experimentation through percentage splitter model training and testing option by using the same algorithm. Therefore, experiment one conducting through J48 decision tree classifier algorithm is extended by using 80% of the dataset for training and 20% of dataset for testing the model performance. The following table 4.16 shows the confusion matrix output of the current experiment.

Table 4.16:-Confusion matrix of J48 decision tree with percentage split

Actual class	Classified class							
	Bac-Pneu	Vi-Pneu	Bro-Pneu	Myc-Pneu	VA-Pneu	Asp-Pneu	Fun-Pneu	Str-Pneu
Bac-Pneu	115	0	1	0	0	1	0	0
Vi-Pneu	1	39	0	0	0	0	0	0
Bro-Pneu	0	0	4	0	0	0	0	0
Myc-Pneu	0	0	0	9	0	0	0	0
VA-Pneu	0	0	0	0	5	0	0	0
Asp-Pneu	2	0	0	0	0	61	0	0
Fun-Pneu	0	0	0	0	0	0	10	0
Str-Pneu	0	0	0	0	0	0	0	10

As the above confusion matrix output shown, in among 8 classes of pneumonia J48 algorithm incorrectly classified 2, 1 and 2 cases of Bacterial-Pneumonia, Viral-Pneumonia and Aspiration-pneumonia into other classes. Moreover the algorithm correctly classified Broncho-Pneumonia, VA-Pneumonia, Mycoplasma-Pneumonia, Fungal-Pneumonia, and Streptococcus-pneumonia cases in to their actual classes without error. On the above experiments the algorithm generate 40 number of leaves and 75 tree sizes which is similar to the previous experiment results.

After building the classification model the next task was also evaluation the specific experiments results. Since, the following table shows J48 classifier algorithm model results that was built through 80% for training and 20% of the data for training.

Table 4.17:-Summary of J48 decision tree classifier experiment result

Model characteristics	Experiment results
Accuracy	98.06%
Time taken	0.05 seconds
AV.TPR (%)	0.981
AV.FPR (%)	0.011
AV.PR	0.981
AV.RR	0.981
AV.ROC	0.987
F-Measure	0.981
CCI	253
ICI	5

As the above table 4.17 shown, from the total 258 test instance the algorithm correctly classified 253(98.06%) in to their actual classes, whereas only 5 (1.93%) of instance are incorrectly classified. The algorithm also takes 0.05 second in order to build the model. The ROC performance measure indicator show the algorithm perform 98.7%.

As clearly shown on the above two experiments conducted through J48 decision tree algorithm, it generate 40 number of leaves and 75 nodes. This show there are some outliers in the data which needs to be detected and removed through pruned the tree. However, the researcher was perform the above two experiments J48 pruned algorithm parameters in WEKA data mining tool. Reducing size of tree and leaves in order easily understand generated rule or tree is crucial. Since the researcher perform the following experiment by changing default “minimum number of objects “or minNumObj= 5, 10 and 15. Since, the good results obtained from minNumObj= 15 decision tree using J48 algorithm conducted through 10-fold cross validation option is presented in this study.

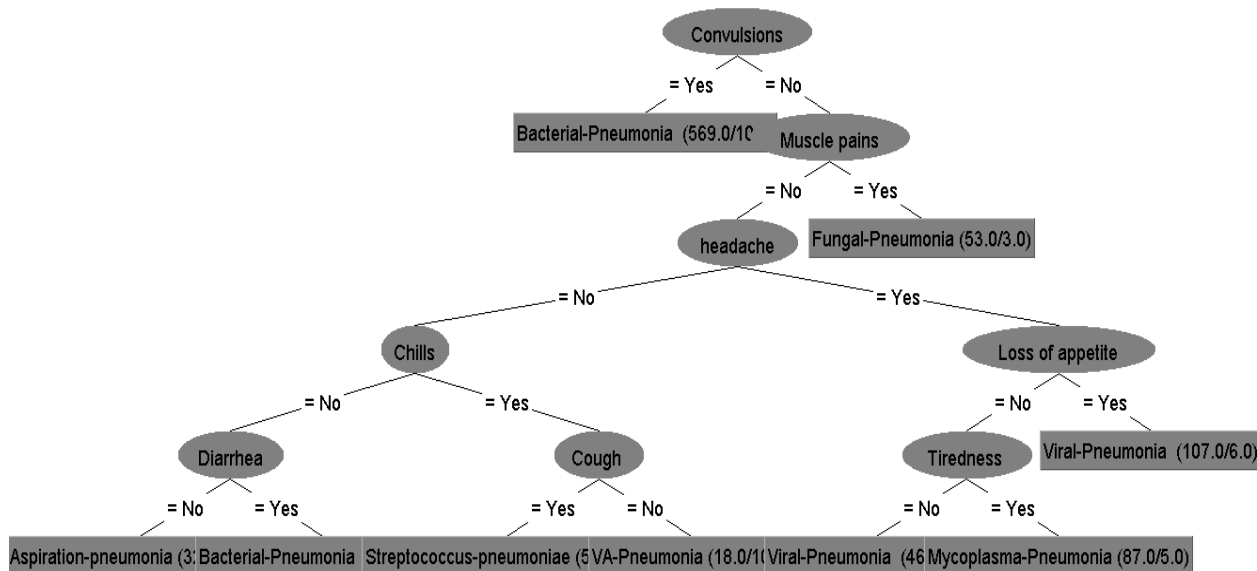


Figure 4.3:-A tree generated from J48 pruned tree

The above tree generated from J48 pruned tree by configuring minNumObj or setting minimum number objects as 15. Accordingly, the generated tree were reduced into 9 number of leaves and 17 size of tree. However, the above tree structure is simple for understanding by domain experts, where the algorithm accuracy reduced into 93.41% which less as compared with the first experiment accuracy. Therefore, the researcher takes the first experiment conducted through default parameters which minNumObj=2 because of the accuracy of the model.

4.4.1.2 Experimentation Two

In this particular experiment, PART rule classifier algorithm were applied in order to build classifier model. Therefore this experimentation is conducted through WEKA default or 10-fold cross validation model training and testing option.

Table 4.18:-Confusion matrix of PART with 10-fold cross validation

Actual class	Classified class							
	Bac-Pneu	Vi-Pneu	Bro-Pneu	Myc-Pneu	VA-Pneu	Asp-Pneu	Fun-Pneu	Str-Pneu
Bac-Pneu	582	4	1	0	0	3	0	2
Vi-Pneu	2	158	1	0	0	0	2	0

Bro- Pneu	0	0	27	0	0	0	0	0
Myc- Pneu	0	0	0	82	0	0	0	0
VA- Pneu	0	0	0	0	14	0	0	0
Asp- Pneu	4	3	0	0	1	342	1	0
Fun- Pneu	0	0	0	0	0	1	50	0
Str- Pneu	0	0	0	0	0	0	0	47

As the above confusion matrix output clearly showed, PART rule classifier algorithm correctly classified Broncho-Pneumonia, Mycoplasma-Pneumonia, VA-Pneumonia and Streptococcus-pneumonia cases without any error, whereas on the remaining pneumonia classes there are some cases which misclassified into other classes. Moreover, among 1291 total instances, 1266(98.06%) cases are correctly classified into their classes, whereas the remaining 25(1.93%) of cases were misclassified.

On KDD data mining process model once build the classifier model the next task was evaluation of the model performance. Since, the following table 4.19 showed the result of detailed accuracy the current experiment conducted through PRAT classifier with 10-fold cross validation.

Table 4.19:- Summary of PART classifier experiment result for 10-fold cross validation

Model characteristics	Experiment results
Accuracy	98.06%
Time taken	0.08 seconds
AV.TPR (%)	0.981
AV.FPR (%)	0.006
AV.PR	0.981
AV.RR	0.981
AV.ROC	0.989
F-Measure	0.981
CCI	1266
ICI	25

The model which build through PRAT rule classifier algorithm in this specific experiment scored 98.06% accuracy and it takes 0.08 seconds to build the model. Moreover, the algorithm generate 28 rules and ROC Area measure indicter the algorithm perform 98.9%.

Another experiment were also conducted using PART classifier algorithm by changing 10-fold cross validation setting into percentage spilt model training and testing option. Accordingly, this

experiment is conducted through using 80% of the data for training the model and 20% of the data for testing the trained model performance.

Table 4.20:- Summary of PART classifier experiment result with percentage split

Actual class	Classified class							
	Bac-Pneu	Vi-Pneu	Bro-Pneu	Myc-Pneu	VA-Pneu	Asp-Pneu	Fun-Pneu	Str-Pneu
Bac-Pneu	116	0	1	0	0	0	0	0
Vi-Pneu	2	39	1	0	0	0	1	0
Bro-Pneu	0	0	4	0	0	0	0	0
Myc-Pneu	0	0	0	9	0	0	0	0
VA-Pneu	0	0	0	0	5	0	0	0
Asp-Pneu	2	0	0	0	0	61	0	0
Fun-Pneu	0	0	0	0	0	0	10	0
Str-Pneu	0	0	0	0	0	0	0	10

As shown on the above confusion matrix result of the above experiment, PART rule classifier algorithm correctly classified all Broncho-Pneumonia, Mycoplasma-Pneumonia, VA-Pneumonia, Fungal-Pneumonia and Streptococcus-pneumoniae cases as it is. However, some of Bacterial-Pneumonia, Viral-Pneumonia and Aspiration-pneumonia cases incorrectly classified into other classes.

The PART classifier model also evaluated in the following table 4.21 experiment results depicted from WEKA data mining tool.

Table 4.21:- Summary of PART classifier experiment result

Model characteristics	Experiment results
Accuracy	98.44%
Time taken	0.26 seconds
AV.TPR (%)	0.984
AV.FPR (%)	0.007
AV.PR	0.986
AV.RR	0.984
AV.ROC	0.99
F-Measure	0.985
CCI	254
ICI	4

As clearly shown the experiment result of PART rule classifier algorithm, the algorithm correctly 254(98.44%) of cases and incorrectly classified only 4(1.55%) cases into other classes. The

algorithm also takes 0.26 seconds in order to build which high but had a better accuracy as compared with the pervious experiment conducted through the same algorithm. ROC performance measure of the algorithm also shown better than all previously conducted experiments which is 99% performance.

4.4.1.3 Experimentation Three

In this particular experiment, Naïve Bayes probabilistic classifier algorithm were applied in order to build classifier model. Similar with pervious experiments in this experimentation 10-fold cross validation and 80/20 percentage split model training and testing option were applied. Hence, in this particular default 10-fold cross validation option were applied.

Table 4.22:-Confusion matrix of Naïve Bayes with 10-fold cross validation

Actual class	Classified class							
	Bac-Pneu	Vi-Pneu	Bro-Pneu	Myc-Pneu	VA-Pneu	Asp-Pneu	Fun-Pneu	Str-Pneu
Bac-Pneu	580	4	3	0	0	5	0	0
Vi-Pneu	12	140	1	4	0	5	1	0
Bro-Pneu	0	0	15	0	0	12	0	0
Myc-Pneu	0	0	0	82	0	0	0	0
VA-Pneu	0	0	0	0	8	0	0	6
Asp-Pneu	14	10	18	5	0	268	0	0
Fun-Pneu	0	0	0	0	0	0	51	0
Str-Pneu	0	0	0	0	0	0	0	47

As clearly shown on the above table 4.22 confusion matrix output of Naïve Bayes classifier algorithm, all Mycoplasma-Pneumonia, VA-Pneumonia , Fungal-Pneumonia and Streptococcus-pneumoniae test cases or instances are correctly classified. Whereas, on the remaining classes some cases or instances were misclassified in to other classes by the algorithm.

Table 4.23:-Summary of Naïve Bayes classifier experiment result

Model characteristics	Experiment results
Accuracy	92.25%
Time taken	0.01 seconds
AV.TPR (%)	0.923
AV.FPR (%)	0.025
AV.PR	0.927
AV.RR	0.923
AV.ROC	0.994

F-Measure	0.923
CCI	1191
ICI	100

The specific experiment conducted through Naïve Bayes classifier with 10-fold cross validation model training and testing option, the algorithm scored 92.25% accuracy. Among 1291(100%) cases, 1191(92.25%) of them are correctly classified, whereas the remaining 100(7.74%) cases were misclassified. Moreover, the algorithm takes 0.01 second to build the classifier model. Furthermore, the ROC performance measure the algorithm perform 99.4%.

The experiment conducted through Naïve Bayes classifier algorithm also extended by changing model training and testing option in to percentage splitter in which 80% of the data were used for model training and 20% of the data for testing.

Table 4.24:-Confusion matrix of Naïve Bayes with percentage splitter

Actual class	Classified class							
	Bac-Pneu	Vi-Pneu	Bro-Pneu	Myc-Pneu	VA-Pneu	Asp-Pneu	Fun-Pneu	Str-Pneu
Bac-Pneu	116	0	0	0	0	1	0	0
Vi-Pneu	3	36	0	0	0	0	1	0
Bro-Pneu	0	0	2	0	0	2	0	0
Myc-Pneu	0	0	0	9	0	0	0	0
VA-Pneu	0	0	0	0	2	0	0	3
Asp-Pneu	2	3	1	1	0	56	0	0
Fun-Pneu	0	0	0	0	0	0	10	0
Str-Pneu	0	0	0	0	0	0	0	10

As clearly shown on the above table 4.24 confusion matrix output of Naïve Bayes classifier algorithm, three classes of pneumonia namely Mycoplasma-Pneumonia, Fungal-Pneumonia and Streptococcus-pneumoniae all test cases are correctly classified as it is. However, on the remaining five classes of pneumonia some cases were misclassified in other classes.

After build the model the next step or phase on KDD data mining process model was evaluation of the specific model building algorithm performance and accuracy. Therefore the following table 4.25 shows the experiment result and evaluation of Naïve Bayes classifier conducted with percentage splitter option. Summary

Table 4.25:- Summary of Naïve Bayes classifier experiment result

Model characteristics	Experiment results
Accuracy	93.41%
Time taken	0.01 seconds
AV.TPR (%)	0.934
AV.FPR (%)	0.023
AV.PR	0.936
AV.RR	0.934
AV.ROC	0.994
F-Measure	0.931
CCI	241
ICI	17

In the above particular experiment, Naïve Bayes classifier algorithm scored 93.41% accuracy and takes 0.01 seconds to build the classifier model. Moreover, from 258 test cases, the algorithm correctly classified 241(93.41%) cases or instances into their actual classes, whereas the remaining 17(6.58%) cases were misclassified.

4.4.2 Comparison of classification algorithms

After all of building and testing the model, there is a need of choose of the best classifier model in order to acquire more accurate cases form the data mining. Since, the J48 decision tree, PART rule and Naïve Bayes classifier model are compared with some criteria's. Therefore, the researcher evaluate and compare the above three algorithm based on objective criteria's which means the aim acquiring cases were mainly used to take more accurate cases into cases based systems. Hence, overall accuracy, correctly and incorrectly classified cases, time taken to build the model, precision and recall rate criteria's or metrics were used to choose the best model in in this study.

Table 4.26:-Comparison of classification algorithms

		Classification Algorithms		
Objective evaluation criteria's	Experiment setting	J48	PART	Naïve Bayes
Overall Accuracy	10-fold cross validation	97.98%	98.06%	92.25%
	80/20%	98.06%	98.44%	93.41%
Correctly classified cases	10-fold cross validation	1265	1266	1191
	80/20%	253	254	241
Incorrectly classified	10-fold cross	26	25	100

cases	validation			
	80/20%	5	4	17
Time taken	10-fold cross validation	0.11 sec	0.08 sec	0.01 sec
	80/20%	0.05 sec	0.26 sec	0.01 sec
TP rate	10-fold cross validation	0.98	0.981	0.923
	80/20%	0.981	0.984	0.934
FP rate	10-fold cross validation	0.007	0.006	0.025
	80/20%	0.011	0.007	0.023
Precision	10-fold cross validation	0.98	0.981	0.927
	80/20%	0.981	0.986	0.936
Recall	10-fold cross validation	0.98	0.981	0.923
	80/20%	0.981	0.984	0.934

The above table 4.26 shows a comparison of classification algorithms that was used in classifier model building experimentations. From overall accuracy of the system, PART rule classifier algorithm conducted through 80/20% percentage splitter model training and testing option had 98.44% accuracy which is better than the remaining two algorithms.

Moreover, PART rule classification algorithm perform better in all objective evaluation criteria's than J48 and Naïve Bayes except time taken to build classification model. From time taken to build classifier model criteria, Naïve Bayes classifier algorithm takes only 0.01 seconds which is better than J48 and PART algorithms. Therefore, based on the objective classifier algorithm evaluation criteria's, the researcher decide to used PART rule classifier algorithm model to taken cases for cases based reasoning systems.

Once decide to use PART classifier model, the next task is test of the selected model performance in classifying new instances. Since, the researcher used 258(20% of the data) cases separately using percentage splitter technique before starting the data mining experimentation. Therefore, the researcher used supplied test set model testing method in WEKA data mining tool. PART classifier model perform 97.83% prediction accuracy on separate test dataset. Since, the

following depicted figure is taken from PART classifier model prediction results by click and visualize classifier errors and saving the classifier result into .arff file format.

12: chest pain Nominal	13: headache Nominal	14: sweating Nominal	15: Difficulty swallowing Nominal	16: Nausea Nominal	17: Dizziness Nominal	18: Fussiness Nominal	19: Shortness of breathing Nominal	20: Skin rash Nominal	21: Wheezing Nominal	22: Muscle pains Nominal	23: Convulsions Nominal	24: Place Nominal	25: predictedClassification Nominal	26: Classification Nominal
No	Yes	No	No	No	No	No	No	Yes	No	Yes	No	CAP	Fungal-Pneumonia	Fungal-Pneumonia
Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	No	No	CAP	Aspiration-pneumonia	Aspiration-pneum...
No	No	No	No	No	No	No	No	No	No	No	No	HAP	Streptococcus-pneumoniae	Streptococcus-pn...
No	No	No	No	No	No	No	No	No	Yes	No	Yes	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia
Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	No	No	CAP	Aspiration-pneumonia	Aspiration-pneum...
No	Yes	No	No	No	No	No	No	No	No	No	No	CAP	Mycoplasma-Pneumonia	Mycoplasma-Pneu...
Yes	No	Yes	Yes	No	No	No	Yes	No	Yes	No	No	CAP	Aspiration-pneumonia	Aspiration-pneum...
Yes	Yes	No	No	No	No	No	No	Yes	Yes	No	No	CAP	Mycoplasma-Pneumonia	Mycoplasma-Pneu...
No	Yes	No	No	No	No	No	No	Yes	Yes	No	No	CAP	Mycoplasma-Pneumonia	Mycoplasma-Pneu...
Yes	No	No	No	No	No	No	Yes	No	Yes	No	No	CAP	Aspiration-pneumonia	Aspiration-pneum...
Yes	No	No	No	No	No	No	No	No	Yes	No	Yes	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia
Yes	No	No	No	No	No	No	No	No	No	No	Yes	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia
No	No	No	No	No	No	No	No	No	No	No	Yes	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia
Yes	No	No	No	No	No	No	Yes	No	Yes	No	No	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia
No	Yes	No	No	No	No	Yes	No	No	No	No	No	CAP	Viral-Pneumonia	Viral-Pneumonia
No	No	No	No	No	No	Yes	No	No	No	No	No	CAP	Viral-Pneumonia	Viral-Pneumonia
No	No	No	No	No	No	No	No	No	No	No	Yes	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia
Yes	No	Yes	Yes	No	No	No	No	No	No	No	No	HAP	Aspiration-pneumonia	Aspiration-pneum...
No	No	Yes	Yes	No	No	No	Yes	Yes	Yes	No	No	CAP	Aspiration-pneumonia	Aspiration-pneum...
No	Yes	No	No	No	No	No	No	No	No	No	No	CAP	Viral-Pneumonia	Viral-Pneumonia
No	No	No	No	No	No	No	No	No	No	No	No	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia
Yes	No	Yes	No	No	No	No	No	No	No	No	Yes	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia
No	No	No	No	No	No	No	No	No	No	No	Yes	CAP	Bacterial-Pneumonia	Bacterial-Pneumonia

Figure 4.4:-A sample of test instances using PART classifier algorithm

As we shown on the above figure 4.5, PART classifier algorithm predict instances into their pneumonia classes. For instance, as shown on the above figure the algorithm predict the first instance Fungal-Pneumonia correctly in to Fungal-Pneumonia in which the 25 attribute values showed the new predicted results of the algorithm.

However, the above test cases results only showed the predict classes values which is correct or incorrect, whereas there is need of looking more visualized method in order to take more accurate cases for cases beads system. Since, the researcher used model building and testing by commands on WEKA “Simple CLI” application. Model building and testing through commands

on “Simple CLI application” allowed to visualize individual (each instances) test cases prediction accuracy as correct or incorrect.

In order to build the model and save the model for using while for testing new cases, the researcher used the following commands.

```
java weka.classifiers.rules.PART -C 0.25 -M 2 -split-percentage 80 -t
C:\Users\Dan\Desktop\MYDM\training.arff -d
C:\Users\Dan\Desktop\MYDM\DanielfinalModel.model
```

On the above commands the first line commands shows the algorithm used to build the model which is PART rule classifier algorithm through using 80% of the data for model building and 20% the data for testing. The commands “C 0.25 “refer the confidence factor values and “-M 2” also refer minimum number of instances takes or minimum the algorithm takes minimum number of two instances. Moreover, on the second line of commands –t refer traning the model by accessing the traning data from the specific location. Also, on the last command –d refer devleop the model and save the model on the above specific location as “TrianingModel.model”. Therefore the following depicted figure 4.6 from Simple CLI showed the result of the generated model using the above commands.

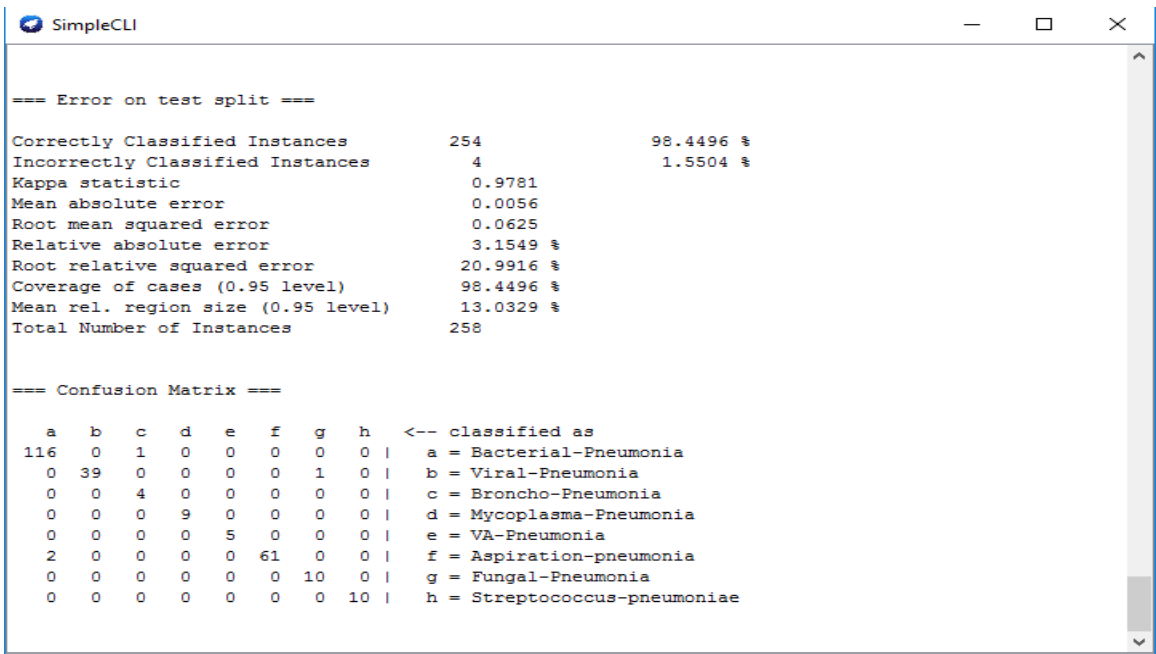


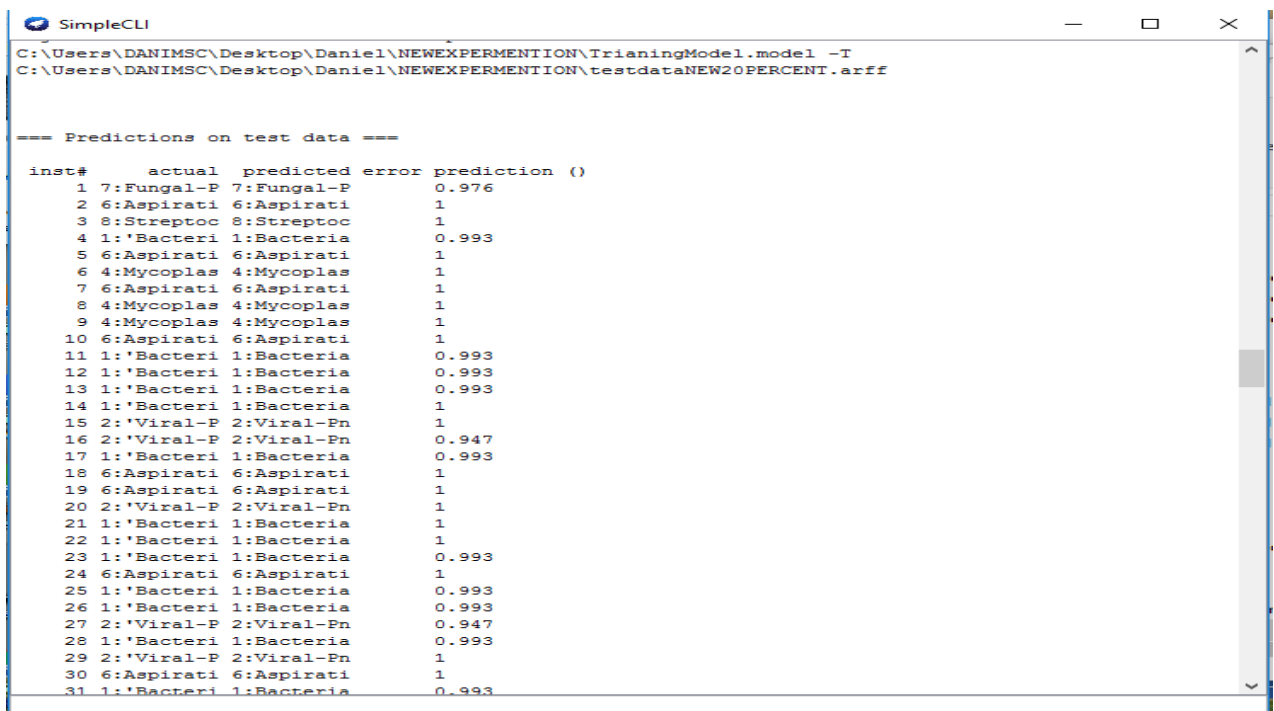
Figure 4.5:-model building result through commands on WEKA “Simple CLI”

After build and save the prediction model the task is testing of new instances or cases by loading the model from the specific location. Since, the researcher used the following commands in order to test new instance, this is mainly allowed to visualize individual instances test accuracy or confidence level whether they are predicted correctly or incorrectly.

```
java weka.classifiers.rules.PART -p 25 -l
C:\Users\Dan\Desktop\MYDM\DanielfinalModel.model
-T C:\Users\Dan\Desktop\MYDM\test.arff
```

On the above first commands, the researcher need to predict(-p) attribute 25 values or type of pneumonia cases using PART rule classifier algorithm by accessing the generated model from the specific location listed on the second commands. Therefore, the on the above line of commands “-L” refer to load the model from the above specific location which is called “TrianingModel.model”. After load the model, the command “-T” used to run test of the model by accessing the test data from the specified location.

After running the above commands, the result of new instances were displayed under WEKA “Simple CLI” command interface. Therefore, the following figure 4.5 showed predict new instances result of PART classifier model.



```
SimpleCLI
C:\Users\DANIMSC\Desktop\Daniel\NEWEXPERMENTION\TrianingModel.model -T
C:\Users\DANIMSC\Desktop\Daniel\NEWEXPERMENTION\testdataNEW20PERCENT.arff

---- Predictions on test data ----

inst#      actual   predicted error prediction ()
1 7:Fungal-P 7:Fungal-P 0.976
2 6:Aspirati 6:Aspirati 1
3 8:Streptoc 8:Streptoc 1
4 1:'Bacteri 1:Bacteria 0.993
5 6:Aspirati 6:Aspirati 1
6 4:Mycoplas 4:Mycoplas 1
7 6:Aspirati 6:Aspirati 1
8 4:Mycoplas 4:Mycoplas 1
9 4:Mycoplas 4:Mycoplas 1
10 6:Aspirati 6:Aspirati 1
11 1:'Bacteri 1:Bacteria 0.993
12 1:'Bacteri 1:Bacteria 0.993
13 1:'Bacteri 1:Bacteria 0.993
14 1:'Bacteri 1:Bacteria 1
15 2:'Viral-P 2:Viral-Pn 1
16 2:'Viral-P 2:Viral-Pn 0.947
17 1:'Bacteri 1:Bacteria 0.993
18 6:Aspirati 6:Aspirati 1
19 6:Aspirati 6:Aspirati 1
20 2:'Viral-P 2:Viral-Pn 1
21 1:'Bacteri 1:Bacteria 1
22 1:'Bacteri 1:Bacteria 1
23 1:'Bacteri 1:Bacteria 0.993
24 6:Aspirati 6:Aspirati 1
25 1:'Bacteri 1:Bacteria 0.993
26 1:'Bacteri 1:Bacteria 0.993
27 2:'Viral-P 2:Viral-Pn 0.947
28 1:'Bacteri 1:Bacteria 0.993
29 2:'Viral-P 2:Viral-Pn 1
30 6:Aspirati 6:Aspirati 1
31 1:'Bacteri 1:Bacteria 0.993
```

Figure 4.6:-Sample CLI prediction results on test data using PART

As clearly shows figure 4.6, the PART classifier algorithm predict test instance with different pneumonia classes wither correctly or incorrectly with their corresponding confidence level or individual instances prediction accuracy. For example, from the above results the instances actual class was “Bacteria-Pneumonia” and the algorithm predict as “Bacteria-Pneumonia” correctly with 0.99 confidence level or 99 % accuracy.

Once we have obtained the test cases with their corresponding confidence factor it is possible to take more accurate cases into case based system which is mainly enhance the quality system as compared as manually collecting and inserting cases. Therefore, as clearly shown on the above experimentation, the selected algorithm PART algorithm has 98.44% accuracy which shows most instance or cases are averagely above 98.44%. since the researcher taken test cases which has more than 99.0% accuracy in order to enhance the quality of the study. So among 323 cases, 291 cases were used for the prototype development.

In general, data mining techniques are showed its significance to acquire filtered knowledge/cases for effective implementation of cases based reasoning system. In order obtained better cases from large amount of pneumonia patient cases/records, the researcher was used PART rule classifier algorithm because of its better accuracy on classifying different pneumonia cases. This allowed to selected higher accuracy cases, to select few but representative cases and to reduced size of cases stored on case-based database for effective development and implementation case-based reasoning system.

CHAPTER FIVE

SYSTEM DEVELOPMENT AND EVALUATION

5.1. Architecture of the Prototype System

Figure 5.1 depicts the architecture of the CBRSDTPUFYC system that shows the application of data mining for constructing cases used for designing case based reasoning system for diagnosis and treatment of pneumonia under-five year children.

Once the case based system is designed, new query (problem) is entered through user interface, and the system searches the best matching cases form the case base by using similarity measurement. If relevant cases are found within the case base, then the system rank the relevant retrieved cases based on their global similarity. Next, the system proposes a solution. The proposed solution (solved solution) can be derived directly from a retrieved case that matches exactly or partially to the problem of the new case. But, using the proposed solutions directly may have a risk. Therefore, the user of the system should make an adaptation by altering the differences between the proposed case and the new case. In addition to adaptation, case inconsistencies are revised if the retrieved case is not the same as the new case. Finally, the revised solution is retained by incorporating it into the existing case-base for future problem solving.

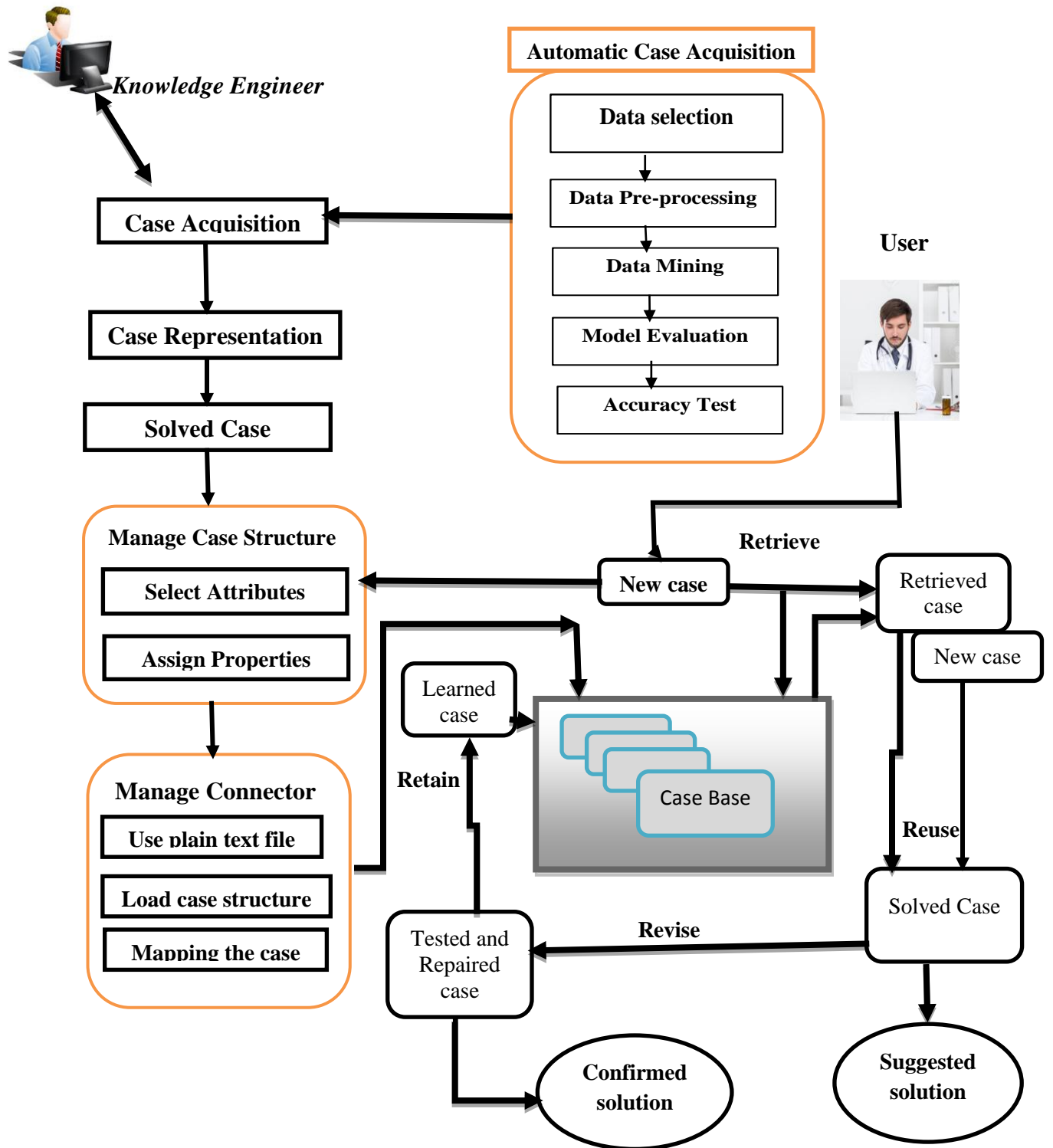


Figure 5.1:-Architecture of the CBRSDTPUFYC system

5.2. Case Based Reasoning System for CBRSDTPUFYC

The development of a CBR application involves number of steps, such as collecting cases and background knowledge, modeling, case representation, defining an accurate similarity measure, implementing retrieval functionality, and implementing user interfaces (Armin & Thomas, 2008). In this study, the researcher used the main feature of jCOLIBRI to develop the actual prototype. As Recio-Garcia, Diaz-Agudo and González-Calero (2008) presented jCOLIBRI has been constructed as a core module to offer the basic functionality for developing CBR application. Implementing a CBR application from scratch remains a time consuming software engineering process and requires a lot of specific experience beyond pure programming skills (Armin & Thomas, 2008). Therefore, using jCOLIBRI CBR framework minimizes the effort to develop an application by using other programming languages.

To run JCOLIBRI for the first time, click on the JCOLIBRI.bat file and it becomes ready for usage as shown in the following figure 5.2

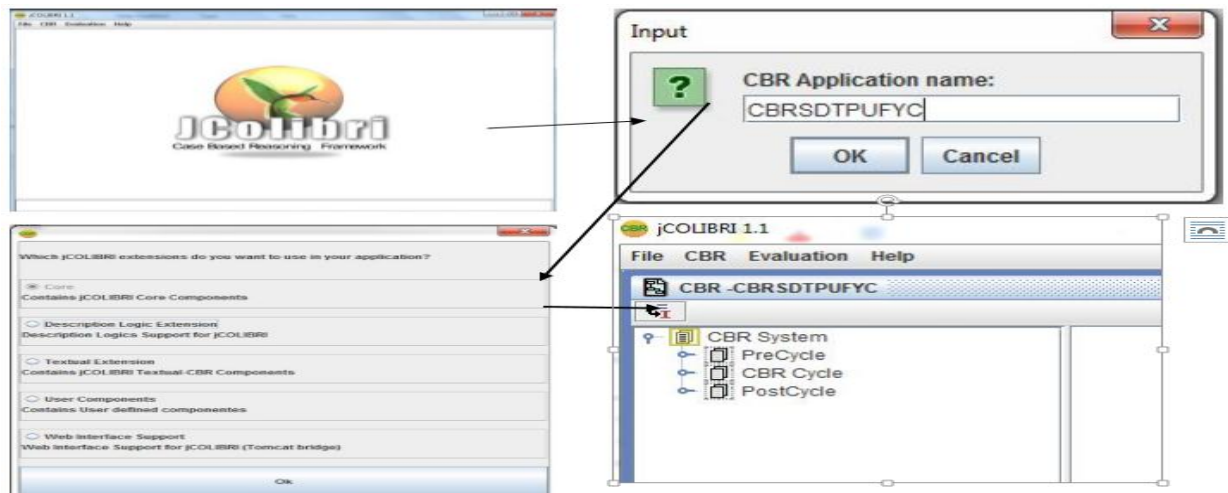


Figure 5.2:- Main and Configuration Window of JCOLIBRI

In this study, the development of CBR system for diagnosis and treatment of pneumonia under five year children is divided in the following sub processes which enable to achieve the objectives of the research.

5.2.1. Building a Case-Base

In order to build a CBR system, collecting the appropriate pneumonia patient cases is the first stage. After that, the collected cases can be stored in representational method. So, the researcher collected pneumonia for under-five year children patient cases from JUSH and used Automatic knowledge Acquisition technique using data mining tool WEKA. As discussed in Chapter 4, the test instances that scores more than 99 % of accuracy are used for CBR system. The acquired cases are used to build a case-based diagnosis and treatment of pneumonia under-five year children. All the acquired cases are stored as plaintext files in a feature-value representation format, as a result the .ARFF results of the test cases in to .txt format. The case base is presented as a plain text comprising of n columns representing case attributes (A1, A2, A3, ..., An) and each m rows representing individual cases (C1, C2, C3, ..., Cm), in which each attribute has a sequence of possible values associated to each column attribute $A = \{V1, V2, V3, \dots, Vk\}$. The reason for representing cases using feature-value representation is that this approach supports nearest neighbor retrieval algorithm and it represents cases in an easy way (Tamir, A et al, 2017; Mulugeta & Million, 2019)

5.2.2. Case Representation

The case representation is made in a way that easily fit to JCOLIBRI. Designing of such a case structure helps easily define the features available in the cases and used to measure the similarity between existing cases and the new case (query). Hence, the overall application of this research is to retrieve similar cases from the case base that can guide future reasoning, problem solving and also transforming a solution retrieved in a solution appropriate to the current problems. The simplest way to represent a case is by using feature-value representation to make efficient retrieval process. The reason for representing the cases using feature-value representation is that this approach supports nearest neighbor retrieval algorithm and it represents cases in an easy way. This is done through case indexing process. Indexing refers to assigning indices to cases for retrieval and comparison of a query to the case base (Mulugeta & Million, 2019)

5.2.3. Managing the Case Structure in JCOLIBRI

The acquired cases are saved in plaintext file format. As illustrated in figure 5.2 below: cases were constructed with selected attributes with their mapped attribute name, data type, weight, and similarity measures. In this study, twenty four description attributes and two solution

attributes were used and each attributes are mapped to its corresponding weight, similarity methods and data types. Here also there are two alternatives for adding attributes add simple and add compound. Add simple is used to add a single attribute whereas add compound is used to add attribute with a sub attributes inside the major attribute. In this stage all the selected description and solution attributes are managed properly as shown below.

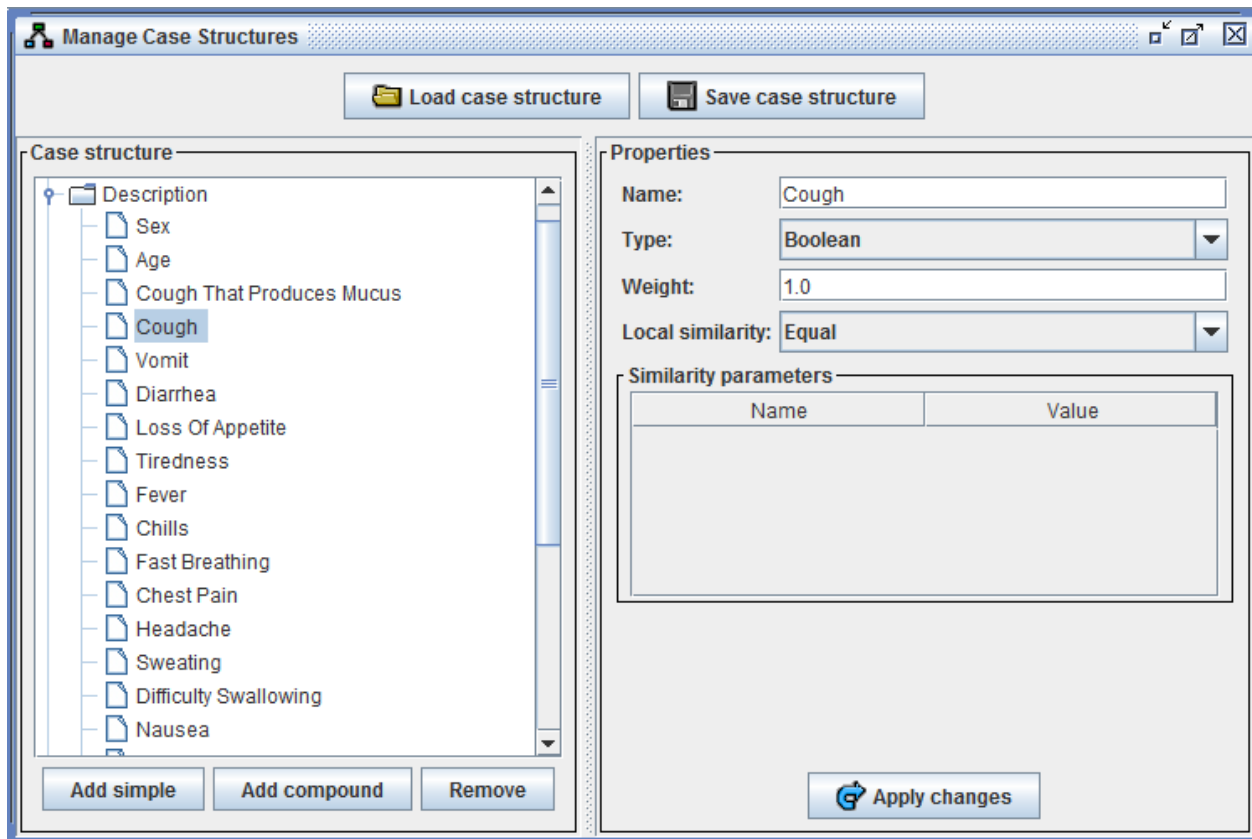


Figure 5.3:- Managing Case Structure In jCOLIBRI

5.2.4. Description of CBRSDTPUFYC Case Attributes

Defining the case structure in JCOLIBRI is done using a simple manage case structure window. It is very easy to define the case structure with JCOLIBRI. Because it is simple to add attributes in the description of case structure and set properties of attributes of metadata of attributes. Metadata of attributes are including weight of attribute, data type of attribute and similarity function during configuration of case structures, JCOLIBRI creates codes automatically and saved in XML file format.

Most significant attributes are set by declaring higher weight as compared to other weights. Based on attribute selection task using information gain attribute evaluator results Convulsions, headache, difficulty swallowing, skin rash, chills, sweating, vomit, fast breathing, diarrhea, loss of appetite, muscle pains, wheezing, place, tiredness, shortness of breath, cough that produces mucus, age and chest pain have more weight than other attributes. For building CBRSDTPUFYC the weights value for the attributes comes from attribute selection using information gain attribute evaluation and domain expert's feedback on the results. Based on the result, Convulsions, headache, fast breathing, tiredness, shortness of breath, cough that produces mucus, age, chest pain, cough, fever, diagnosis result and recommended treatment attributes got a weight of 1.0. And also the other remaining attributes weights are assigned by discussing with domain experts. The description of case attributes regarding attributes name, data type, weight value and local and global similarities are shown in the following table (Table 5.1).

Table 5.1:- List of attributes and description

Description of Attributes				
No	Attributes Name	Data Type	Weight	Local Similarity
1	Sex	String	0.7	Equal
2	Age	String	1.0	Equal
3	Cough That Produces Mucus	Boolean	1.0	Equal
4	Cough	Boolean	1.0	Equal
5	Vomit	Boolean	0.9	Equal
6	Diarrhea	Boolean	0.9	Equal
7	Loss Of Appetite	Boolean	0.9	Equal
8	Tiredness	Boolean	1.0	Equal
9	Fever	String	1.0	Equal
10	Chills	Boolean	0.8	Equal
11	Fast Breathing	Boolean	1.0	Equal
12	Chest Pain	Boolean	1.0	Equal
13	Headache	Boolean	1.0	Equal
14	Sweating	Boolean	0.8	Equal
15	Difficulty Swallowing	Boolean	0.8	Equal
16	Nausea	Boolean	0.7	Equal

17	Dizziness	Boolean	0.7	Equal
18	Fussiness	Boolean	0.7	Equal
19	Shortness Of Breath	Boolean	1.0	Equal
20	Skin Rash	Boolean	0.7	Equal
21	Wheezing	Boolean	0.8	Equal
22	Muscle Pains	Boolean	0.8	Equal
23	Convulsions	Boolean	1.0	Equal
24	Place	String	0.8	Equal
Solution Attributes				
25	Diagnosis Result	String	1.0	Equal
26	Recommended Treatment	String	1.0	Equal

5.2.5. Managing Connectors

After configuring the case structure in jCOLIBRI, CBR systems must access the stored cases in an efficient way. JCOLIBRI, splits the problem of case base management in two separate although related concerns: persistency mechanisms through connectors and in-memory organization. A connector is an object which has the ability to access and retrieve cases from a specific case persistency when given the case structure and gives those cases to the CBR system in a standardized way. Therefore connectors provide an abstraction mechanism that allows users to load cases from different storage sources in a transparent way. In this regard figure 5.4, JCOLIBRI includes connectors that work with plain text files, relational databases and Description Logics systems.

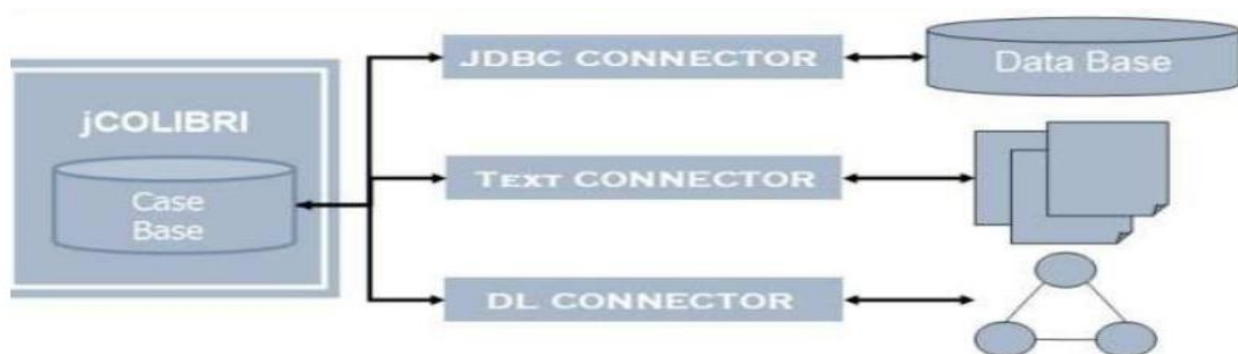


Figure 5.4:- JCOLIBRI Connector Schema

For the implementation CBRSDTPUFYC prototype, the researcher used plaintext connector because pneumonia cases are stored in plaintext file format after Data mining model evaluation. The generated text file is loaded through managing connector task by specifying the path of case structure and path of text file. And also the punctuation ‘Comma’ is used as a separator between the values of the attributes. All the attributes of a case should be mapped. This is connector’s responsibility to retrieve data from case base and return it back to GUI. Finally save the connector in xml file format.

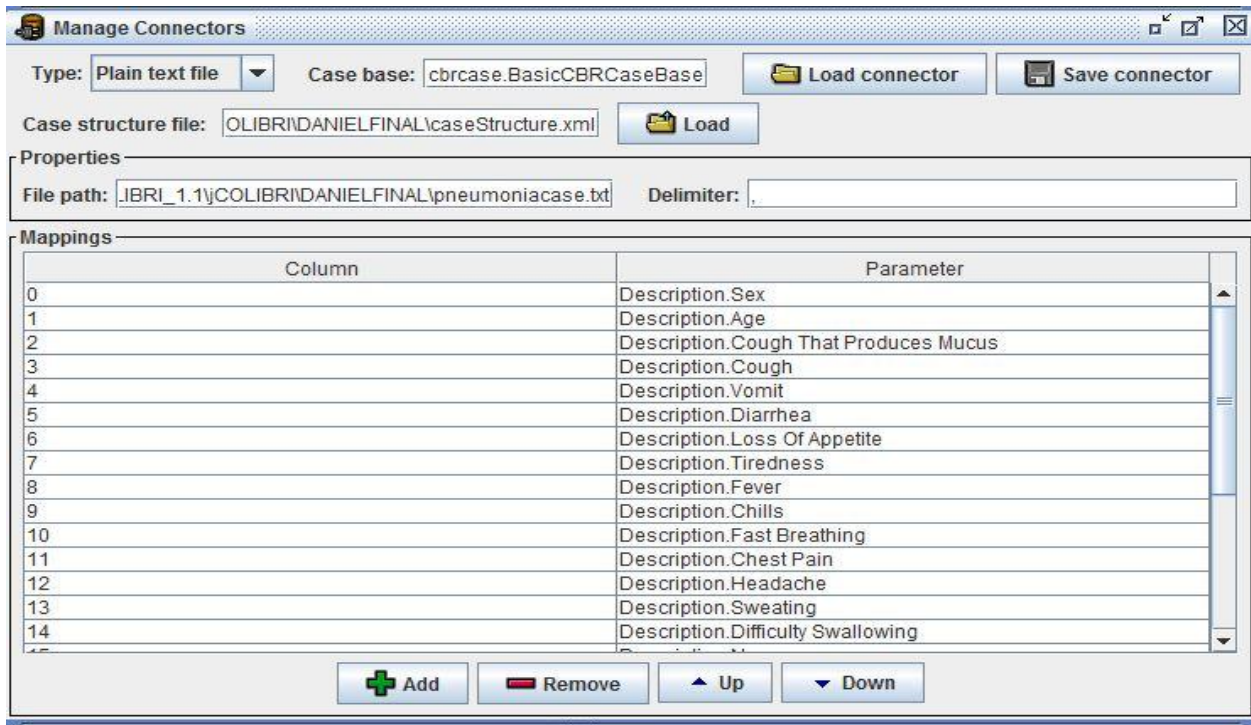


Figure 5.5:- Managing Connectors window

5.2.6. Managing Tasks and Methods

5.2.6.1. Managing Tasks

After designing the case structure and managing the connectors, the next task is loading the cases and performs all activities until the cases are stored on the persistence layer. In development of the prototype CBRSDTPUFYC, the researcher used three core tasks a Pre Cycle that loads cases, the CBR Cycle and a Post Cycle that stores cases into the persistence layer.

PreCycle task executes before the main CBR cycle. Its task is to get all the cases in case base. Therefore, it is necessary to define path of connector in its subtask. There is only one subtask

called obtain case task and it is used retrieve data from case base before the execution of the main CBR cycle.

Main CBR cycle is the main task of CBR cycle and it also has sub tasks. The developer has to give path of case structure in it. It knows number of case attributes that are available. It is called obtain query task. In addition to obtain query task, there are other significant tasks under the main CBR cycle. These are retrieve tasks, reuse tasks, revise task and retain tasks.

Obtain query task: used to obtain the query that contains the description of the problem and used to retrieve the most similar cases. Clinical Signs and symptoms for a suspect of pneumonia marked on the displayed query window for the process of diagnosis and treatment of pneumonia for under-five age children.

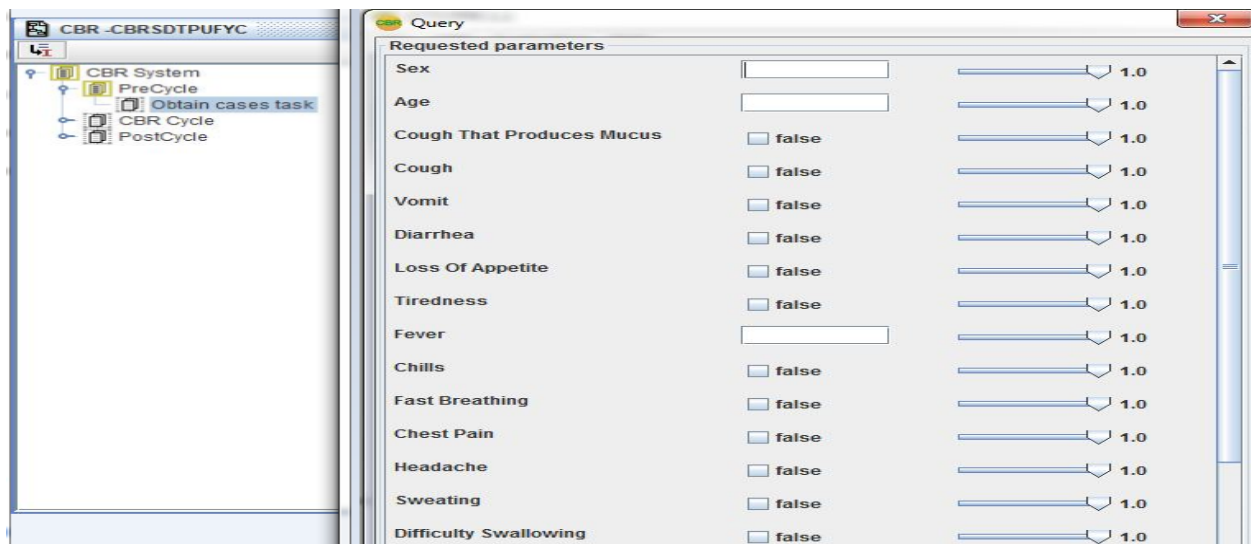


Figure 5.6:- Window for Obtaining query task

Retrieve tasks used to retrieve case from the stored case base. Retrieve tasks are three subtasks. The subtasks include select working cases task, compute similarity task and select the best case. Select working case task selects cases from case base and stores them into current context. Compute similarity sub task compute similarity of the stored cases with the case entered by the user using the query window. Select best case shows the best matched of case(s) after computing the similarity of stored cases against the new case. It means that the number of best matched case(s) is shown to the user depending on the method used and the threshold.

Reuse tasks enable to reuse previously stored cases. It has three subtasks. These subtasks are: prepare cases for adaptation task, automatic reuse task and reuse task. Prepare cases for adaptation task select cases from case base and stores them into context for the requirement of the new solution. Here also specifying the path of case structure in this method is needed. Atomic reuse task should be resolved by reuse resolution method.

Revise task is the evaluation stage about the selected solution in reuse phase. After selecting the most similar cases from the retrieved results, the solution for the problem should be confirmed and validated by domain experts before the solution is stored in a case base for future use.

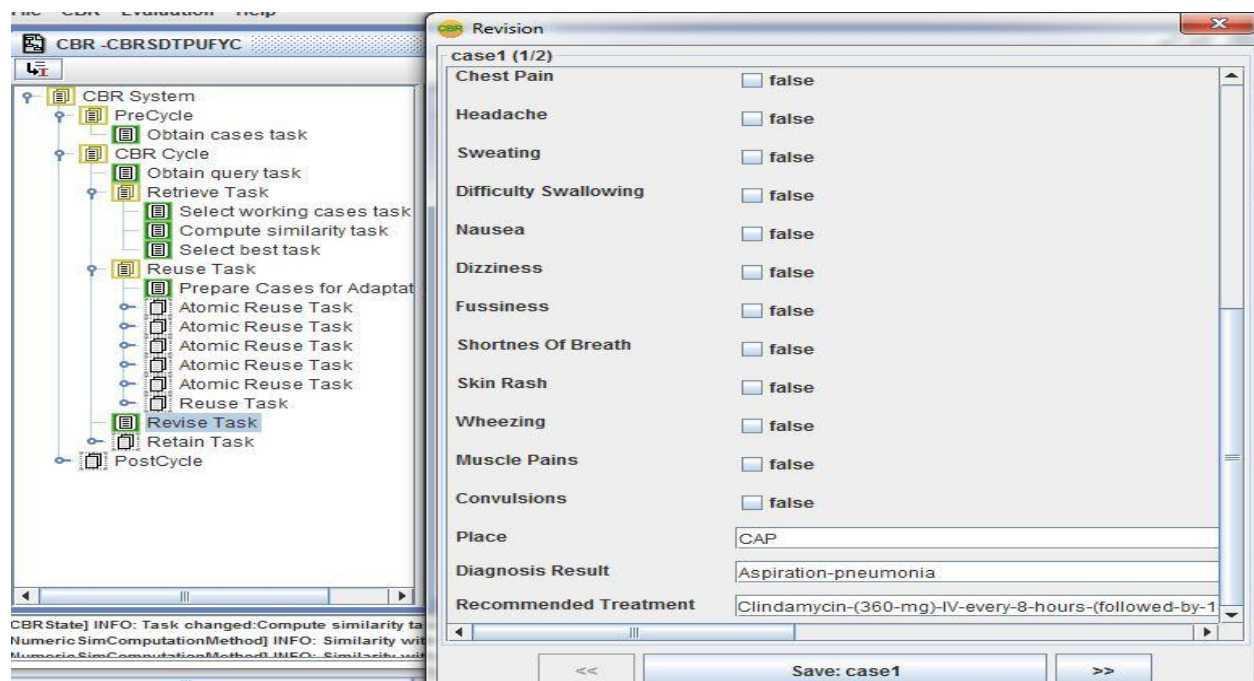


Figure 5.7:- Window of Case Revision

Retain tasks also used to CBR case retention on a persistence layer. It has also its own subtasks like select cases to store task and store cases task. First, the revised case has to be selected to be stored. Secondly, the selected case be stored into the case base and prepared to be stored on the persistence layer.

The last task in managing tasks in JCOLIBRI is PostCycle. PostCycle task have only one sub task called close connectors task which is usually executed after the main CBR cycle. Its main task is to close a connection between case base and GUI.

Case Similarity, Matching and Ranking

The primary goal of CBR system is to retrieve best similar cases to the query from case base using some similarity assessment of heuristic functions. The similarity function involves in computing the similarity between the stored cases in the case base and the query, and selects nearest similar cases to the query. Therefore, jCOLIBRI uses the nearest neighbor algorithm as a cases retrieval technique. Nearest neighbor algorithm used to measure the similarity between the stored and the new queries, and return the search results within their ranked order. For each attribute in the query and case, local similarity function measures the similarity between two simple attribute values. Based on the matching weighted sum features from those simple attributes, the similarity score between the queries and stored cases for each simple attribute is assigned.

Finally, the average score (global similarity) of each attribute between the case and the query are computed and the result is assigned to the object (the similarity between the stored case and the query). The maximum degree of similarity among the retrieved cases is displayed according to their ranked order.

5.2.6.2. Managing Methods

The method library stores classes that actually resolve the task. These classes can resolve the CBR cycle using in programming or using GUI. All tasks that are mentioned above should have their own methods to be assigned in order to achieve the tasks goal. The following are lists of methods which are used to solve tasks for this CBRSDTPUFYC application.

LoadCaseBaseMethod: This method returns the whole available cases from the case base to designer. This method use connector to retrieve case base.

ConfigureQueryMethod: Displays the GUI window in which the user can enter query to retrieve cases from the case base. It uses case structures as input parameters

SelectAllMethod: Selects working cases from case base and store them into current context. It allows displaying all the available cases from the case base to the result window.

SelectBestCaseMethod: This method selects the best similar case among the displayed cases, by prioritizing the similarity results.

NumeriSimilarityComputationalMethod: this is used to calculate similarity between the query and cases that are stored in the case base. It uses nearest neighbor similarity for the computation.

NumericProportionMethod: it is the sub method of reuse task which involve in computing numeric proportion between the description attributes and solution attributes.

ManualRevisionMethod: it allows the users to modify cases in the query window as they need. User can change case according to his/her will especially in the revise task.

RetainChooserMethod: This method allows the user to choose the method. Chosen method will store case base. User can choose that he/she want this method to store in case base.

StoreCasesMethod: This method used to stores cases into Case base.

CloseConnectorMethod: Closes the connector by saving the case permanently to the case base.

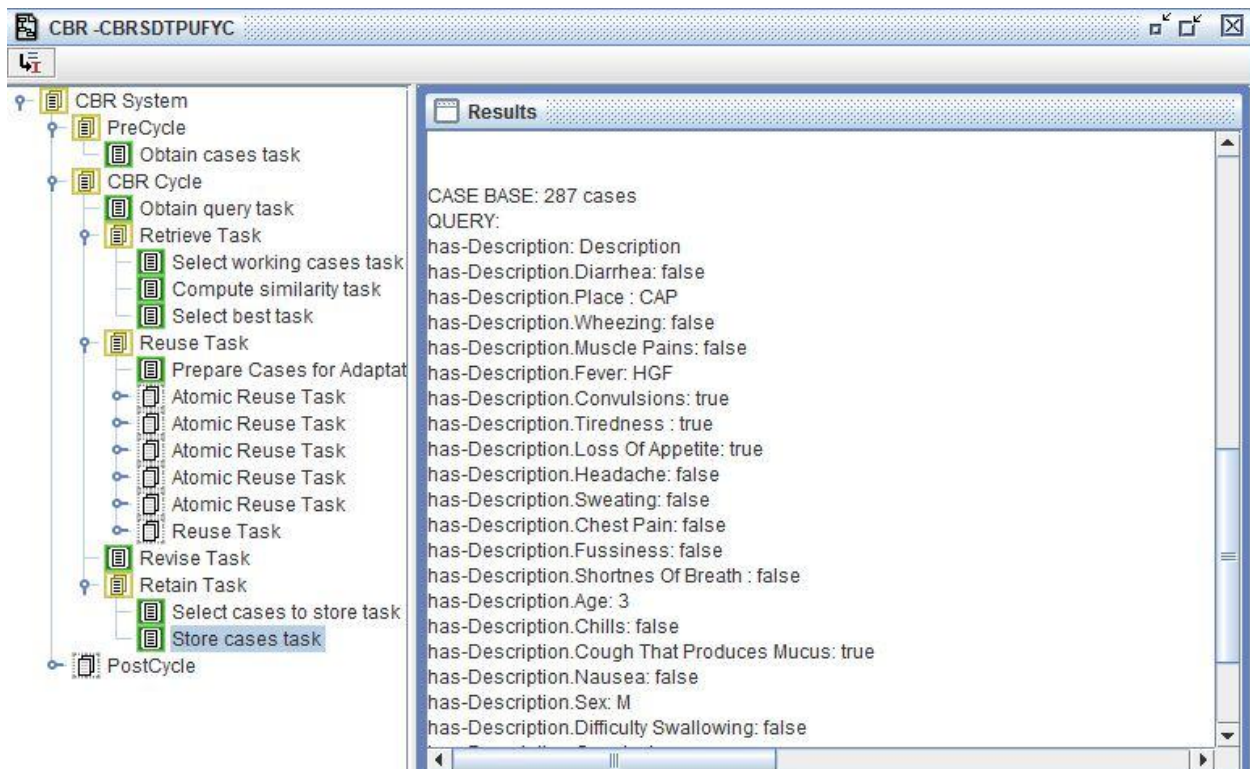


Figure 5.8:-Tasks and methods configuration

In this chapter, the researcher discussed how the prototype CBRSDTPUFYC works by using jCOLIBRI. All the four CBR cycle components such as; retrieve, reuse, revise and retain were

applied and discussed in the developed prototype system. The developed CBRSDTUPFYC achieved the requirement of the study in giving advice for diagnosis and treatment of pneumonia for under-five year children.

Once the CBR system is ready based on the knowledge acquired using data mining classification technique, the next task is checking its performance to determine whether the prototype system meets the level of accuracy as required. It confirms whether the right prototype is developed. To this end the researcher evaluate the system performance using test cases and also user acceptance testing.

5.3. System Evaluation

Testing and performance evaluation is an important issue for every intelligent system (i.e. Knowledge Based System). This chapter presents users' evaluation of the developed system as well as the performance evaluation of the system. Therefore, to realize the objectives of this study, acquire patient cases from JUSH by using data mining tool WEKA to build the case base. For this research a total of 291 cases which is scored more than 99% accuracy during Data Mining test are used for CBR system. The testing method used for evaluating the performance of the prototype system was made by using the parameters precision, recall and F-measure. These three parameters were used in order to measure the accuracy of the prototype system. In addition to this, user acceptance testing of the prototype is performed by domain experts.

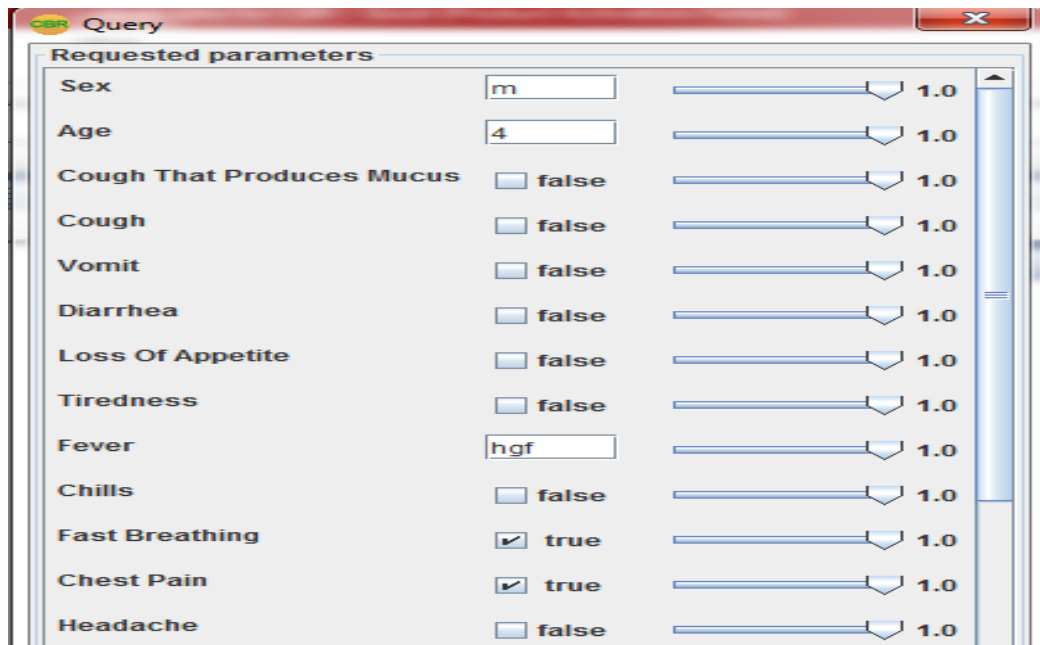
5.3.1. Testing the CBR Cycles and Evaluating the Performance of the prototype

To check the validity and performance of the CBR system to domain experts, the functionality of CBR cycles and effectiveness of the prototype should be tested with selected test cases. The effectiveness of the prototype is measured with recall, precision and F- measure using test cases. In addition to that, the performance of the system was evaluated from the user's side with user acceptance testing. With user acceptance testing, potential users of the system rate the applicability of the system in their day to day activities.

5.3.1.1. Retrieve and Reuse Evaluation

To retrieve any suggestion from the system, first the user should initiate the system by parsing query which is the short expression of the users need. As figure 5.8 illustrate dominant attributes

are shown with their defined weight and on this interface the user register what he/she want to ask the system.



The screenshot shows a window titled "Query" with a sub-header "Requested parameters". It contains a list of 13 medical symptoms, each with an input field, a checkbox, a slider, and a weight value of 1.0. The input fields contain the following values: Sex (m), Age (4), Cough That Produces Mucus (false), Cough (false), Vomit (false), Diarrhea (false), Loss Of Appetite (false), Tiredness (false), Fever (ngf), Chills (false), Fast Breathing (checked true), Chest Pain (checked true), and Headache (false).

Parameter	Input Field	Checkbox	Slider	Weight
Sex	m	<input type="checkbox"/>	100%	1.0
Age	4	<input type="checkbox"/>	100%	1.0
Cough That Produces Mucus	false	<input type="checkbox"/>	100%	1.0
Cough	false	<input type="checkbox"/>	100%	1.0
Vomit	false	<input type="checkbox"/>	100%	1.0
Diarrhea	false	<input type="checkbox"/>	100%	1.0
Loss Of Appetite	false	<input type="checkbox"/>	100%	1.0
Tiredness	false	<input type="checkbox"/>	100%	1.0
Fever	ngf	<input type="checkbox"/>	100%	1.0
Chills	false	<input type="checkbox"/>	100%	1.0
Fast Breathing	checked true	<input checked="" type="checkbox"/>	100%	1.0
Chest Pain	checked true	<input checked="" type="checkbox"/>	100%	1.0
Headache	false	<input type="checkbox"/>	100%	1.0

Figure 5.9:-Query Interface

After registering the query, the user needs solution. So retrieve similar cases to the new case from previously solved cases is followed by the reuse of similar solutions. In this research retrieval of cases is performed using the nearest neighbor retrieval algorithm because the implementation tool JCOLIBRI uses this algorithm. During retrieval, similar cases are retrieved to the new case with appropriate ranking. After that the user of the system can use the solution of the retrieved cases in a way that can fit to the problem at hand.

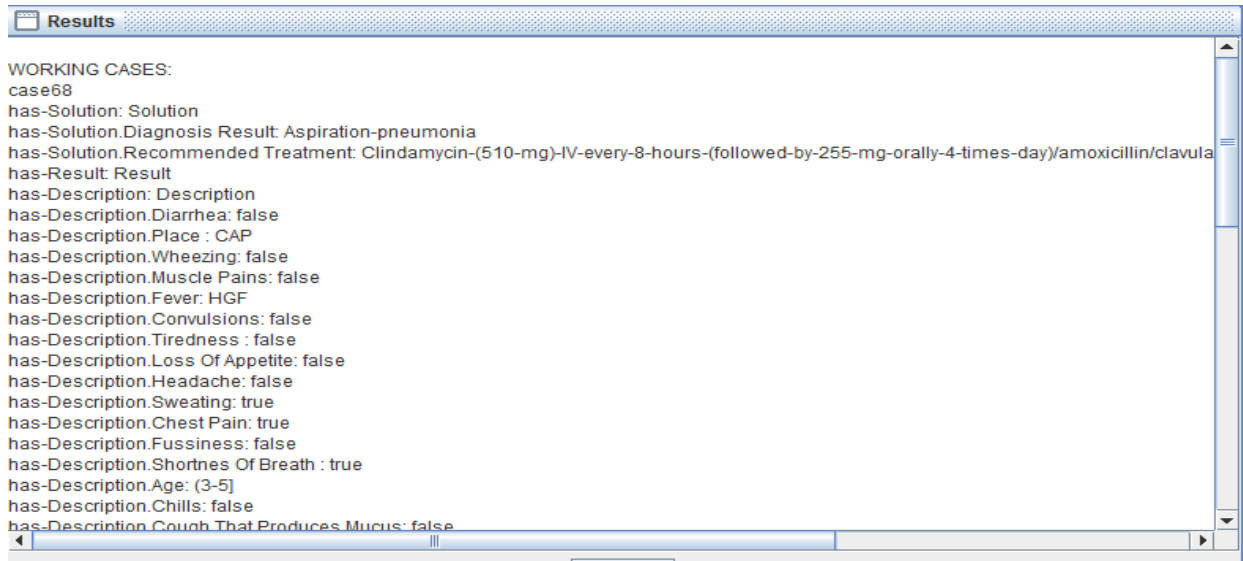


Figure 5.10:-Retrieved Solution case

The statistical analysis evaluation uses 291 pneumonia for under-five year children cases that have been collected from JUSH using data mining acquisition techniques. In this study, the effectiveness of the retrieval process of the CBRSDTPUFYC is measured by using recall, precision and F- measure (harmonic mean of recall and precision). While recall is the ability of the retrieval system to retrieve all relevant cases to a given new problem (query) from the case base. On the other hand, precision is the proportion of retrieved cases that are relevant to a given query.

To conduct the evaluation, for each test case the relevant pneumonia for under-five year children 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 case base to each of the test cases. The domain expert uses the value of disease class and Recommended Treatment or solution attributes of the pneumonia for under-five year children 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, recall, precision and F-measure are calculated.

Precision and recall can be calculated with the following formulas.

$$\mathbf{Recall} = \frac{\text{Number of relevant cases retrieved}}{\text{Total number of relevant cases in a case base}}$$

$$\mathbf{Precision} = \frac{\text{Number of relevant case retrieved}}{\text{Total number of retrieved cases}}$$

$$\mathbf{F_{Measure}} = \frac{2(\text{Recall} * \text{Precision})}{\text{Recall} + \text{precision}}$$

Table 5.2:-Relevant cases assigned by domain experts for sample test cases

Test case	Relevant case from case base
Case1	case111, case114, case143, case146, case43, case61, case83, case122, case130, case148, case123, case191
Case2	Case76, case123, case59, case97, case110, case112, case122
Case 3	Case4, case41, case62, case174, case191, case199, case231, case250, case38, case144
Case4	Case39, case53, case93, case247, case284, case38
Case 5	Case7, case282, case283, case128, case8, case122
Case 6	Case137, case150
Case 7	Case166, case290, case83

Once the relevant cases are identified and assigned to the test cases the next step is calculating the recall, precision and F-measure value to measure the performance of the CBR system. Previous researchers used (1.0, 0.8) threshold according to Henok (2011) and Mekedes (2018) indicated in his research, there is no standard threshold for the degree of similarity that has been used for retrieving relevant cases in CBR. Different CBR researchers use different case similarity threshold. Henok (2011) and Mekedes (2018) used a threshold level of [1.0, 0.8) i.e. this means cases with global similarity score greater than 80% are retrieved. In this regard the researcher uses a threshold value between 0.8 to 1.0 means that cases with a similarity degree of at least 0.8 to 1.0 are the most relevant cases for the user query. In this manner, the following table 5.3 constructed for calculating the Recall, Precision and F-measure of the system.

Table 5.3:-Performance Measurement of CBRSDTPUFYC using Precision, Recall and F-measure

Test cases	Relevant cases suggested by domain experts	Relevant cases retrieved by the system	Total retrieved system cases by the system	Recall	Precision	F-measure
Test case1	12	11	13	0.92	0.85	0.88
Test case2	7	7	7	1.00	1.00	1.00
Test case3	10	10	11	1.00	0.91	0.95
Test case4	6	5	6	0.83	0.83	0.83
Test case5	6	6	7	1.00	0.86	0.93
Test case6	2	2	2	1.00	1.00	1.00
Test case7	3	3	4	1.00	0.75	0.86
Average				0.96	0.89	0.9236

As shown on table 5.3, the system retrieved relevant cases with an average precision of 89%. Although, the average precision value score is best, but, few number of cases retrieved are not relevant within the expected cases, because different case types registered in case base of this system has common symptoms when diagnosing the patients with pneumonia to classify the diseases the have into different types of pneumonia. That is why the system doesn't score 100% precision.

The average recall of the designed prototype is 96%. The other evaluation score is F-measure which is the harmonic mean of precision and recall. The system scores an average of 92.36% this also showed good performance of the developed system. Generally, precision, recall and F-measure average values shown us the average performance of the system as good and could be used to assist health professionals for diagnosis and treatment of pneumonia under-five year children.

5.3.2.2. Case Revision and Solution Adaptation Testing

In medical diagnosis adaptation is a commonly required task. Since this research primary objective is developing CBRSDTPUFYC, adaptation is essential. The purpose testing adaptation of solutions is to evaluate the system's capacity to reuse cases from the case base. Initially the system load case bases at the PreCycle stage and then selects working cases from the case base and stores the cases in to current context at the retrieval stage. The next stage is reusing the cases that are loaded in the working memory. If there is no difference between a current case (query)

and the retrieved exact cases, the adaptation is null and the retrieved case is used without adaptation. When the previous solution is not fully reasonable in the current problem, only few modifications are required to fit the current situation. This issue is a serious issue especially in medical diagnosis because of the corresponding risks. Therefore, the adaptation stage requires experienced domain expert knowledge about how differences in problems of previous case and the current situation are occurred. So, it is up to the domain experts to reuse the retrieved cases to solve the new case rather than the system by itself derives solution. Hence, the adaptation stage of CBRSDTPUFYC is left to the users of the system by comparing specified parameters of the retrieved and current case to modify the solution in a way that can fit to the problem at hand.

In general, the adaptation process of CBRSDTPUFYC is effective as the case features of the previous and new case have similar or less inconsistency attribute values. On the other hand, no adaptation process can be performed as the attribute values of the previous and new cases have more dissimilar or totally different from the previous cases. However, often a direct application of an uncertain solution is impossible due to the corresponding risks, especially in medical diagnosis systems. Therefore, the adaptation has to be performed manually by a human domain expert as shown in the figure 5.11 below.

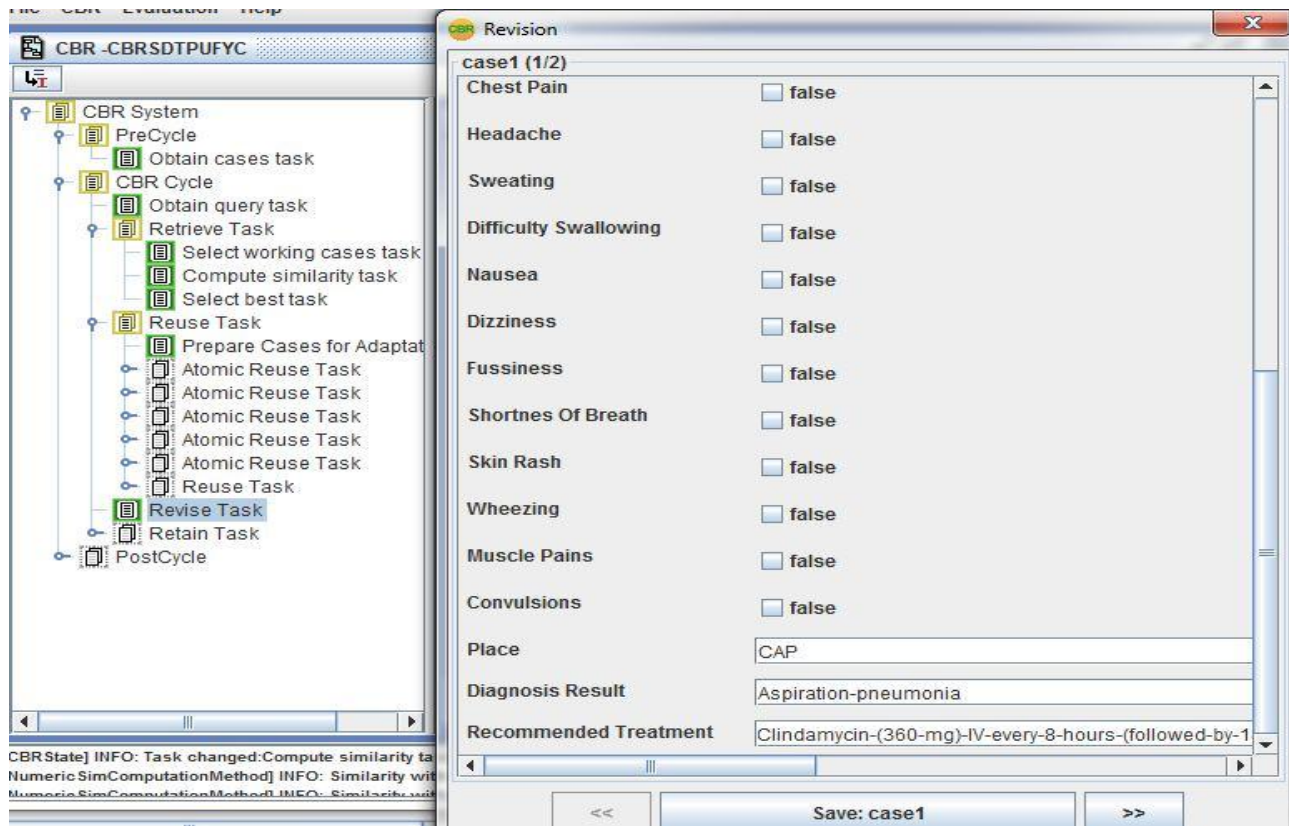


Figure 5.11:-Revision Interface

Lastly, the user has to retain or store the revised case in CBRSDTUPFYC which is an important step in storing new cases which would be used for future diagnosis. The revised case in CBRSDTUPFYC, which doesn't affect the original case, after it is confirmed by the domain experts, will be retained or stored in the case base for future use. As depicted on figure 5.12, retaining cases after revision is possible.

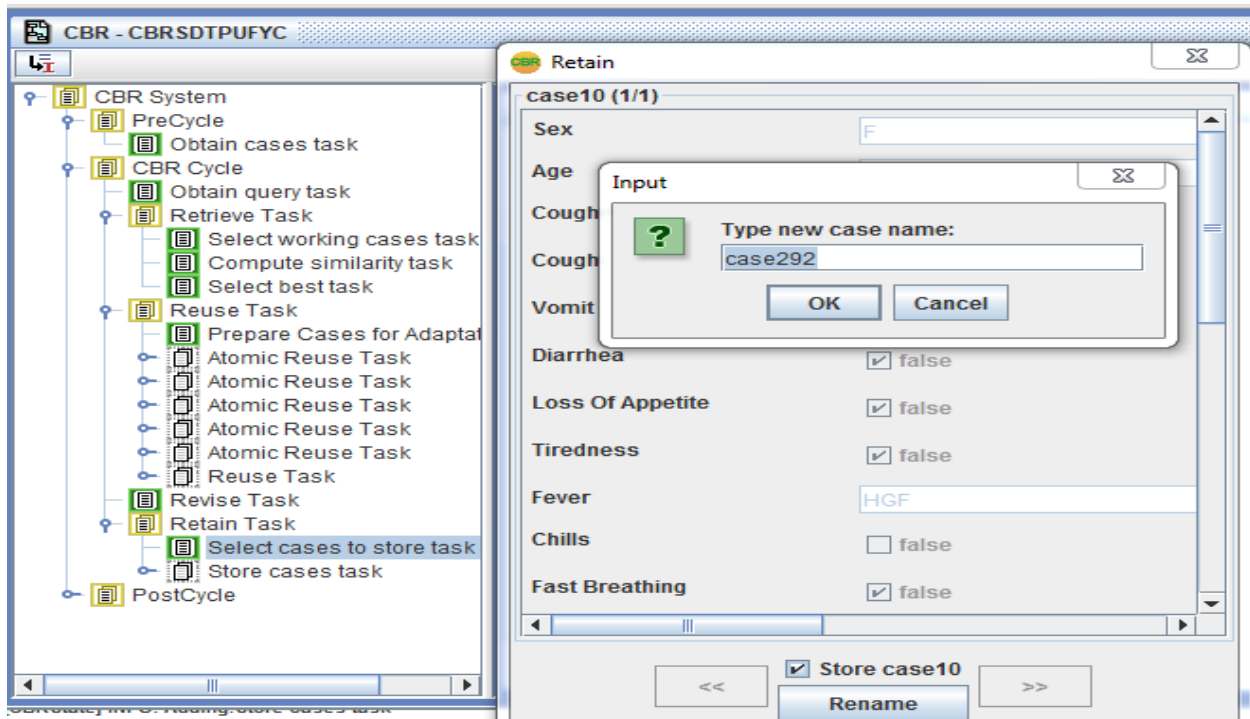


Figure 5.12:-Retaining the revised case

Finally, the user confirms that the CBR learned one additional case and stored in to the case base as shown on figure 5.13.

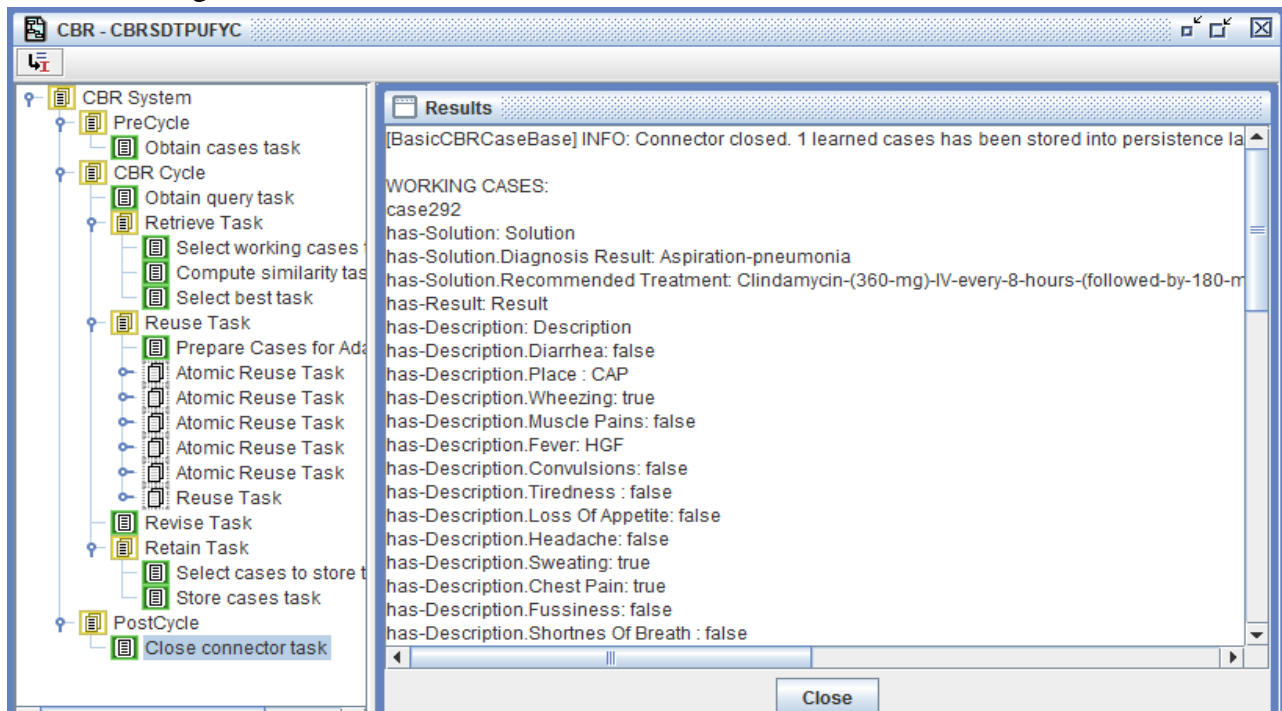


Figure 5.13:-Interface of learned cases

5.3.2. User Acceptance Testing

User acceptance testing is performed in a real situation at JUSH. During testing the user's acceptance the applicability of the prototype is evaluated by potential users of the system. To achieve the goal of user acceptance testing six health professionals (four nurses and two health extension workers) identified and selected purposively from Jimma University specialized hospital. During the system development these domain expert were actively participated in the different stage of knowledge acquisition, prototype development, consulting on the content of knowledge and provide the necessary feedback.

During testing experts are requested to rank each parameter from poor to excellent by assigning value for poor=1, fair=2, good=3, very good=4, excellent= 5.

Table 5.3:-User Acceptance testing from domain experts

No	Evaluation Criteria	Performance Value						
		1	2	3	4	5	Average	Percent (100%)
1	Is the Prototype system ease to use				2	4	4.67	93.4
2	Is the Prototype system adequate and clear for decision support?				1	5	4.83	96.6
3	Relevance of attributes in representing the pneumonia for under-five year children case			1		5	4.67	93.4
4	Fitness of the final solution to the problem at hand				1	5	4.83	96.6
5	Relevance of the retrieved case in the diagnosis and treatment of Pneumonia under-five year children				1	5	4.83	96.6
6	Efficiency of the system in time				2	4	4.67	93.4
7	Is the prototype system user interface interactive?		1	1	1	3	4.00	80
8	Rate the significance of the system in the domain area					6	5.00	100

Total Average	4.69	94
----------------------	-------------	-----------

As indicated in table 5.3, for the first question concerning the prototype system ease to use 33% of the respondents rated the criterion as very good. 67% of the respondent's rate as excellent, the overall average performance score for this evaluation measure is 93%. For the second parameter which was rate adequacy and decision support, 17% respondent rated as very good. 83% of the respondent's rate as excellent, overall average performance score for this evaluation measure is 96.6%. 17% of the respondents rate the relevance of attributes in representing pneumonia for under-five year children cases rate as good whereas 83% of the respondent's rate as excellent, for this measure the overall average performance score is 93%. In the case of fitness of the final retrieved solution to the new problem at hand around 17% of the respondents rate the prototype is very good whereas only 83% of the respondent's rate as excellent, overall average performance score is 96.6%. In the same way, respondents rated "relevance of the retrieved cases in the diagnosis and treatment of pneumonia for under-five year children" as excellent and very good with 83%, and 17% score respectively, overall average performance for this parameter is 96.6%. 33% of the respondents rate the system as very good and 67% of the respondent's rate as excellent in terms efficiency in time for overall average performance of 93%. For the case of "interactive-ness of the user interface", 50% of the respondent rated the system as excellent and 16% of the respondent rated the system as fair whereas for each scale 17% of the respondents rate as good and very good, for this measure the overall performance score is 80%. Finally, for the question related to significance of the system in the domain area 100% of the domain experts rated excellent with an average performance of 100%

Generally, the user acceptance testing for CBRSDTPUFYC achieved average acceptance of 4.69 out of 5 which accounts 94% that showed the importance and applicability of the prototype system in decision making. So it can be concluded that, CBRSDTPUFYC can be used for supporting decisions in diagnosis and treatment of pneumonia for under-five year children.

5.4. Discussion of Results

In this study the researcher developed a case based system by acquiring cases from patient card and using data mining in order to enhance the quality of the system. Therefore, the CBR prototype was evaluated in terms of recall, precision, F-measure and user acceptance evaluation

techniques. As shown in Table 5.3, the average system performance results evaluating by using recall, precision and F-measure is 96%, 89% and 92% respectively. In addition to this, user acceptance testing performed as depicted on table 6.3. So, 94% of the users or domain experts have accepted the prototype.

As per of the researcher knowledge, no one used data mining techniques to acquire the cases, whereas the researcher used J48, PART and Naïve Bayes data mining classification algorithms in order to enhance the diagnosis capability of the system to enable domain experts to diagnose patients effectively. Since, the researcher compare the developed CBR system with previous studies in the same domain.

Table 5.4:-Comparison of the developed CBR prototype system with previous studies

Author and year	Method and Tool used	Achieved results	User acceptance	Significance
Amelework (2017)	CBR, jCOLIBRI		86%	To support experts for diagnosis of tuberculosis
Mekedes (2018)	CBR, jCOLIBRI		83%	To support experts for diagnosis malnutrition under-five year children
Ermiyas and Hailemicheal (2020)	CBR, jCOLIBRI		Not specified	To support experts for diagnosis of Chronic Kidney Disease
Lucky et al. (2017)	CBR, Not specified		95% system accuracy	To support experts for diagnosis of diagnosis of Bowel Disease
Abebayehu (2015)	RBR	83.33% accuracy on test case	90.40%	Used as for diagnosis and treatment of pneumonia
Zhenjia, et al (2020)	Python			To diagnosis and treatment of Pneumonia
Melquiades and Haile (2019).	RBR, ProLog	87.5%	88% users are satisfied	To solve the shortage of skilled medical experts in the area and the problem of language for diagnosis of treatment of Pneumonia patients

Aiyesha et al(2019)	Not specified	random forest scored 97.64%	Not specified	To support experts for differentiate the diagnosis of Tuberculosis and Pneumonia
Hindayati et al. (2020)	CBR		Not specified	It used to determining a calorie diet per day for each person with similarity values based on case-based.
Bezahegn (2017)	WEKA	PART 96.78%		To support experts for pre- diabetes screening
Desalegn (2017)	Data mining and vb.net RBR	J48=72.3%	76%	To determine choice of contraceptive methods
Kedir (2018)	Data mining and vb.net RBR	J48 =95.1515%,	91.43%	To support experts for diagnosis and treatment of diabetes
Daniel (2020)	Data mining (plus "SimpleCLI" and CBR	PART=98.44%	94%	Mainly used to support medical experts to make more accurate decision for diagnosis and treatment pneumonia

As clearly shown on the above table 5.4, as per of the researcher knowledge there is lack of attempts were made similar with the current work. Abebayehu (2015) works is similar with current work, but the system was developed through RBR which is major difference. Whereas the current work is performed for diagnosis and treatments of eight different types of for under-five year pneumonia, but Abebayehu (2015) works was focused only for two types of pneumonia diseases. In addition to that, the current work scored better performance and user acceptance result as compared with Abebayehu (2015) work due to using of data mining as main knowledge acquisition technique.

Moreover, Amelework (2017), Ermiyas & Hailemicheal (2020), Lucky et al. (2017), Hindayati et al. (2020) and Mekedes (2018) have developed using case-based representation technique to reason out the solution of a particular problem. But, the developed CBRs were not used data mining techniques for acquiring and selecting the most useful cases. Since, the researcher fill this gap by acquiring and selecting most useful pneumonia cases by using data mining techniques and developed the prototype system. As a result, the developed prototype system shows an encouraging result as compared with previous studies.

Furthermore, Kedier (2018), Desalegn (2017) and Bezahegn (2017) used data mining techniques to acquire cases for diagnosis and treatment diabetics, used to determine the choice of contraceptive methods health problems and used developing a predictive model for pre- diabetes screening. However, acquiring cases through using the data mining default experimentation settings had its own limitation means test cases through supplied test set shows the given cases are classified correctly or incorrectly. Which means, to be correct or incorrect prediction the algorithm considers the highest accuracy of the given instance among the specified case. However, diseases diagnosis and treatment issues are the most serious cases which needs more accurate systems. To solve that, the researcher developed and tested cases through commands on WEKA SimpleCLI application, this allowed to show individual test cases accuracy to classify correctly or incorrectly.

In any health problems, any systems developed through scientific or non-scientific way need to be more accurate, since health cases are more serious issues which may led to loss of human life. Since, literature reviewed by the researcher add its contribution in the domain by using data mining techniques, specially the researcher generated the classification model through commands on WEKA “SimpleCLI” which ensure each cases accuracy to classify pneumonia cases in to eight class. As a result, the researcher considered individual pneumonia cases whose accuracy is more than 99% in order to enhance the developed cases based system.

Generally, the current work contributions are stated as follows:

- ✓ The prototype case based system is developed for diagnosis and treatment of pneumonia diseases for under five year children’s.
- ✓ The cases used for CBR system are acquired through data mining techniques by conducted different experimentation with three classification algorithm. As a results PART rule classifier perform better and used as best model.
- ✓ In order to enhance the CBR system performance, the researcher build and test cases through commands on WEKA SimpleCLI application. This enable further researchers used the model at any time for further investigation in addition to allow to take more accurate cases in to CBR system.

In general, this study findings showed the possibilities of using CBR system for diagnosis and treatment of pneumonia diseases for under five year Children’s. The result of the developed CBR

system showed an encouraging result to support medical experts on diagnosis and treatment of pneumonia diseases within the shortage of specialized medical experts in the area.

Domain (medical) experts also provide some useful suggestions and comments on the developed CBR system after conducting user acceptance testing of system. Based on the study findings and domain experts suggestions the researcher forward future works on the recommendation sections of the paper.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

In this study, the researcher major attempt to develop a case base reasoning system for diagnosis and treatment of pneumonia diseases for under five year children's by apply different data mining techniques. These amid to support a shortage medical experts on the area and re-using of past experiences of medical experts for effective cure of different category of the pneumonia disease.

The study was conducted having the main goal of developing a prototype CBR system for diagnosis and treatment of pneumonia under five year children decision making by using manual and data mining techniques that can assist the domain experts. To understand domain problem the researcher conducted interview with medical experts of JUSH. The actual pneumonia diseases cases used for this study has been gathered from JUSH. Therefore, 1614 pneumonia diseases under five year children's case have been collected and used for this study. To achieve the overall objectives of the study, the researcher used design science research design method. A KDD data mining process model was also used for knowledge acquisition through specific experimentation conducted using data mining techniques.

To identify the best prediction model for diagnosis and treatment of pneumonia under five year children, three experimentations were conducted using classification algorithms namely J48 pruned, PART, and Naïve Bayes under 10-fold cross-validation mode and percentage splitter model training and testing option. As a result of those experimentation, PART rule classifier algorithm registered better accuracy which was 98.44% among J48 and Naive Bayes which registered 98.06% and 93.41% overall accuracy respectively. Therefore, PART rule classifier algorithm is selected for better case acquisition and for further use in the development of case base reasoning system.

By using the acquired knowledge the, prototype case bases reasoning system is developed using JCOLIBRI Programming tool. The developed CBR system consisted retrieve, re-use, revise and retain basic tasks. The prototype is finally evaluated in terms of retrieval evaluation test using recall, precision and F-measure assessments were performed. In addition to this user acceptance

testing of the prototype performed in which domain experts were selected and assess the designed prototype based on the criteria's provided for them. From this evaluation test, the average recall, precision and F- measure results 96%, 89% and 92.36 % respectively, is also a promising result to apply CBR in the diagnosis and treatment of pneumonia under five year children. Besides to this, user acceptance testing was performed, so, 94% of the users or domain experts have accepted the prototype.

In general, the result gained form this study was encouraging. It showed that the possibility of diagnosis and treatment of pneumonia diseases for under five year children's through the developed cases based reasoning system by using data mining techniques to classify different types pneumonia diseases .

6.2. Recommendations

The system achieves its objectives by demonstrating the applicability of case-based system by developing a knowledge based system for diagnosis and treatment of pneumonia for under five year children. At the beginning of this research, the researcher set up different specific objectives in harmony with the overall general objective of this study. To this end, all objectives are achieved successfully with some challenges and constraints. Therefore, based on findings of the study the following recommendations were forwarded:

- ❖ Data mining techniques was applied on pneumonia patients' baseline datasets in order to generate cases used for developing prototype of CBRSDTPUFYC. But, pneumonia datasets are manually stored which made preprocessing datasets to difficult. Therefore, designing a data base case record management system would enable effective implementation of CBR through data mining techniques is forwarded for future researchers.
- ❖ Integration of two or more independent prototyping approaches increase the reasoning capability, integrating the case based reasoning system with rule based approaches using data mining knowledge acquisition techniques improve the performance of the system and as well as the reliability of the system and hence its recommend as future research direction

- ❖ Further research needs to be conducted with the inclusion of other important attributes that have significant impact on the diagnosis and treatment pneumonia for under five-year children
- ❖ To develop these KBSs, JCOLIBRI Programming tool was used for implementation, but to make this system more interactive and make life easy for potential users, other visual user interface tools should be used, such as Java.
- ❖ The prototype focused on diagnosis and treatment of pneumonia only under-five year children. In order to give a better dimension, it is better to include all age group.
- ❖ CBRSDTUFYC has used nearest neighbor retrieval which linearly increase the retrieval time, so in the future there is a need to incorporate with other retrieval algorithms such as induction retrieval.

References

- Aaron, K. (2018, September 25). *Medical News Today*. (Daniel, Editor) Retrieved July 27, 2020, from file:///G:/mmm/Bronchopneumonia%20Symptoms,%20causes,%20and%20treatment.htm
- Abdel and AinShams . (2016). Knowledge engineering paradigms in the intelligent medical knowledge-based systems. *4th International Conference on Medical Informatics & Telehealth*. London.
- Abdel-Badeeh, M. (2007, November). Case Based Reasoning Technology for Medical Diagnosis. *Proceedings of world academy of science, engineering and technology*, 25, 9-13.
- Abdi, C. (2016, Feb). An Integration of Prediction Model with Knowledge Base System for Motor Insurance Fraud Detection: The Case of Awash Insurance Company S.C. *Unpublished master thesis*.
- Abdul, F. (2012). A Decision Tree Classification Model for University Admission System. (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, 3, 17-21.
- Abebaw, H. (2014, June). Application of case-based reasoning in legal case management: an experiment with ethiopian labor law cases. *Unpublished masters thesis*.
- Abuka, T. (2017). Prevalence of pneumonia and factors associated among children 2-59 months old in Wondo Genet district, Sidama zone, SNNPR, Ethiopia.
- Adedeji & John. (2002). *Fuzzy engineering expert systems with neural network applications* (Vol. 11). United States of America.: Canada.
- Agnar, A & Enric, P. (1994). Case-Based Reasoning: Foundational Issues, Methodological. *AI Communications*, 7, 39-59.
- Ahmad, N et al. (2011). Evaluating the Success Level of Data Mining Projects Based on CRISP-DM Methodology by a Fuzzy Expert System. *In 2011 3rd International Conference on Electronics Computer Technology*.
- Alec, H and George, B. (1999). Applying case-based reasoning techniques in GIS. *13*, 9-25.
- Alemu, T. (2005). Assessment of radiation exposure of diagnostic x-rays among patients and personnel. *Unpublished mastres thesis*.
- Amelework, F. (2017, October 17). Application of case based reasoning for tuberculosis diagnosis. *Unpublished, (Masters thesis)*.

- Amritpal et al. (2015). Pattern Analysis On Banking Dataset. *International journal of scientific & technology research*, 06(4), 23-30.
- Ana & Manuel. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *IADIS European Conference Data Mining*.
- Andrew, R et al. (2009). Diagnosis of ventilator-associated pneumonia. *Journal of Critical Care*.
- Angel. (2019). Artificial Intelligence in Society.
- Antanassov, A & Antonov, L. (2012). Comparative Analysis of Case-based Reasoning. *Journal of the University of Chemical Technology*, 83-90.
- Armin & Thomas. (2008). Rapid Prototyping of CBR Applications with the Open Source Tool myCBR. *German Research Center for Artificial Intelligence (DFKI) GmbH*.
- Belen D et al. (2007, October 16). Building CBR systems with jCOLIBRI. *Science of Computer Programming*.
- Berhanu, A. (2012). developing A Knowledge based system for coffee disease diagnosis and treatment.
- Bezahegn, Z. (2017). Developing a predictive model for Pre-Diabetics screening by using data mining technology . *Unpublished masters thesis*.
- Carol. (2018, November 28). *WebMD Medical*. Retrieved from <https://www.webmd.com/lung/viral-pneumonia>
- CDC. (2012, December). Fungal pneumonia: a silent epidemic Coccidioidomycosis (valley fever). *National Center for Emerging and Zoonotic Infectious Diseases Division of Foodborne, Waterborne, and Environmental Diseases*.
- Cedars-Sinai. (2020). *Aspiration in Babies and Children*. Retrieved July 23, 2020, from cedars-sinai org: <https://www.cedars-sinai.org/>
- Chapman and Clinton. (2000). CRISPDM 1.0 step-by-step data mining guide. *Technical report, CRISP-DM*.
- Chinelo, I. (2016). *Fundamental of Research Methodology and Data Collection*. University of Nigeria, Nsukka.
- Deshpande, M.P & Thakare, D. (2010). Data mining system and applications. A review. *International Journal of Distributed Parallel Systems*.
- Durairaj & Ranjani. (2013, OCTOBER). Data Mining Applications In Healthcare Sector: A Study. *International journal of scientific & technology research*, Vol, 2(Issue 10), 29-35.

- Durairaj & Ranjani. (2013, Oct). Data Mining Applications In Healthcare Sector: A Study. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 2(10), 29-35.
- Ermiyas,B & Hailemichael, K. (2019, December). Chronic Kidney Disease Diagnosis Model Based on Case Based Reasoning. *International Journal of Engineering and Advanced Technology (IJEAT)*, 9(2).
- Essam, A & AbdEl-Badeeh, S. (2008). Case-based reasoning tools from shells to object-oriented frameworks. *The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution"*. Varna,Bulgaria.
- Fang & Songdong. (2007). Case-Based Reasoning for Logistics Outsourcing Risk Assessment Model. *In Proceedings of International Conference on Enterprise and Management Innovation*.
- Fleck, O. (2019, Feb 27). *Streptococcus pneumoniae (pneumococcus): Overview*. Retrieved Aug 20, 2020, from News-Medical.net: [https://www.news-medical.net/health/Streptococcus-pneumoniae-\(pneumococcus\)-Overview.aspx](https://www.news-medical.net/health/Streptococcus-pneumoniae-(pneumococcus)-Overview.aspx)
- Gedefaw, A et al . (2014). Prevalence of pneumonia among under- five children in Este town and the surrounding rural kebeles, Northwest Ethiopia; A community based cross sectional study. *Science Journal of Public Health*.
- Gemechu, J et al. (2018). Prevalence and associated factors of pediatric emergency mortality at Tikur Anbessa specialized tertiary hospital: a 5 year retrospective case review study. *BMC Pediatrics*.
- Getachaw, W. (2012, June). Application of case-based reasoning for anxiety disorder diagnosis. *Unpublished masters thesis*.
- Getachew, B et al. (2017). Data Mining Attribute Selection Approach for Drought Modelling : A Case Study for Greater Horn of Africa. *international Journal of Data Mining & Knowledge Management Process*, 7(4), 111–126.
- Gonzalo and Oscar. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25:2.
- Graham, R. (2018, September 28). *Healthline*. (L. a. Kim, Editor) Retrieved July 23, 2020, from <https://www.healthline.com/health/mycoplasma-pneumonia>
- Grant, M. (2016, Aug 22). *The definition and classification of pneumonia*. (Keith, Editor) doi:10.1186/s41479-016-0012-z

- Gulavani & Kulkarni. (2009, July-December). A review of knowledge based systems in medical diagnosis. *International Journal of Information Technology and Knowledge Management*, 2, 269-275.
- Hamid T et al. (2015). Using Case-Based Reasoning for Diagnosis in Medical Field. *Journal's URL: <http://www.bepls.com>*, 4.
- Hemant & Limaye. (2011, March 1). Study of Seasonal Variation in Groundwater Quality of Sagar City (India) by Principal Component Analysis. *E-Journal of Chemistry*.
- Henok, B. (2011, JULY). A case based reasoning knowledge based system for hypertension management. *Unpublished masters thesis*.
- Himani & Sunil. (2016, April). A Survey on Decision Tree Algorithms of Classification in Data Mining. *International Journal of Science and Research (IJSR)*, 5(4), 2094-2097.
- Hosseini & Behrouz. (2016). Introducing A Hybrid Data Mining Model to Evaluate Customer Loyalty. *Engineering, Technology & Applied Science Research*, 6, 1235-1240.
- Inderpal, s. (2013). A Review on Knowledge-Based Expert System. *International Journal Of Engineering And Computer Science*, 2(6), 1914- 1918.
- Iqbal & Ashraf . (2006, January). Evaluation of JCOLIBRI. *Msc Thesis*.
- Ismail, k. (2018, Feb 14). Expert System for Diagnosis of Chest Diseases Using Neural Networks.
- Jelena,D et al. (2013). *Neural computing in pharmaceutical products and process development*. Retrieved from Encyclopedia of Health Economics,.
- Jiawei, H. (2006). *Data Mining: Concepts and Techniques* (Second Edition ed.). (Asma, Ed.) San Francisco: Morgan Kaufmann.
- John, C. (2008). Emergency Department Crowding, Part 2—Barriers to Reform and strategies to overcome them.
- JohnsHopkins. (2020). *The Johns Hopkins University*. Retrieved July 23, 2020, from <https://www.hopkinsmedicine.org/health/conditions-and-diseases/pneumonia>
- Kang-Won, L & Gyeong-Chul, K. (1991). Knowledge-Based Expert System in Traffic Signal Control Systems.
- Kassaye, M et al. (2016). Demographic and mortality analysis of hospitalized children at a referral hospital in Addis Ababa, Ethiopia. *BMC Pediatrics*.

- Kenenisa, T. (2018). Prevalence and associated factors of pneumonia among under-five children at public hospitals in jimma zone, South West Of Ethiopia. *Unpublished masters thesis*.
- Krishnamoorthy and Rajeev. (1996). *Artificial Intelligence and Expert Systems for Engineers*. EarthWeb.
- Leake, D. (1996). *The present and future, Book: Case-Based Reasoning: Experiences, lessons, and future directions*,. AAAI Press/MIT Press.
- Manikandan. (2010). Data transformation. *Journal of Pharmacology & Pharmacotherapeutics*, 1(2), 126–127.
- McSherry. (2001). Precision and Recall in Interactive Case-Based Reasoning. *In Case Based Reasoning Research and Development (ICCBR)*, 392-306.
- Mekedes, D. (2018). A case based reasoning system for diagnosis of malnutrition for under-five year children. *unpublished mastes thesis*.
- Merriam et al. (2017). Experiences of nurses on the critical shortage of medical equipment at a rural district hospital in South Africa: a qualitative study. *The Pan African Medical Journal*.
- Mihaela. (2006). On the Use of Data-Mining Techniques in Knowledge-Based Systems. *Economy Informatics*, 21-24.
- Mihaela, K. (2006). On the Use of Data-Mining Techniques in Knowledge-Based Systems. *Economy Informatics*, 21-24.
- Mitch. (2018). Pediatric Workforce Shortages Persist. *Pediatric Workforce Shortage Fact Sheet*.
- Mohamed et al. (2014, June 03). A case based expert system for supporting diagnosis of heart diseases.
- Mohammad et al. (2013, Jun 4). Implementation of Predictive Data Mining Techniques for Identifying Risk Factors of Early AVF Failure in Hemodialysis Patients.
- Morrow,B et al. (2008, April). Guideline for the diagnosis, prevention and treatment of paediatric ventilator-associated pneumonia. 98, 255-268.
- Mu-Jung, H et al. (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications*.
- Mulugeta & Million. (2019, April). A Framework for Personal Computer Hardware Fault Diagnosis and Maintenance. *International Journal of Engineering Science and Computing*, 9(4).

- Mulusew, A. (2014). Knowledge and experience sharing practices among health professionals in hospitals under the Addis Ababa health bureau, Ethiopia. *BMC Health Services Research*.
- Myers, Lawrence, & Tuunanen. (2017). Extending Design Science Research Methodology for a Multicultural World Extending Design Science Research Methodology for a Multicultural World. *Distributed under a Creative Commons Attribution*, 76-98.
- Narina T et al. (2016, June). Case Based Reasoning: A Comparative Analysis of CBR Tools. *International Journal of Innovative Research in Computer and Communication Engineering*, 4,(6).
- Nga, T. (2013). Priority Medicines for Europe and the World "A Public Health Approach to Innovation".
- Nwagu, C et al. (2017, December). Knowledge Discovery in Databases (KDD): An Overview. *International Journal of Computer Science and Information Security (IJCSIS)*, 15, 13-16.
- Omkar, Y. (2014, May). Ranking and Searching of Document with New Innovative Method in Text Mining: First. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 3(5), 406-412.
- Pal, S., and Shiu, K. (2004). Foundation of Soft Cased-Based Reasoning. *Wilely Series on Intelligent Systems*.
- Pascal, R et al. (2017, October 08). Case-based reasoning: potential benefits and limitations for documenting of stories in organizations.
- Peffer, K et al. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45-77.
- Philipp, O et al. (2009). Outline of a Design Science Research Process.
- Poonam, T et al. (2011, July). An Effective Knowledge base system Architecture and issues in representation techniques. *International Journal of Advancements in Technology*.
- Prentzas, J. & Hatzilygeroudis, I. (2007). Categorizing Approaches Combining Rule-based and Case based Reasoning. *Journal Expert Systems*, 97-122.
- Priti S & Rajendra A. (2010). Knowledge-Based Systems for Development. Advanced Knowledge Based System:.. *Model, Application & Research*, 1.
- Radhwan,G et al. (2017, June). Popular Decision Tree Algorithms of Data Mining Techniques: A Review. *International Journal of Computer Science and Mobile Computing*, 6(6), 133 – 142.

- Rainer & Lothar. (2000). Case-based Reasoning for Medical Knowledge-based Systems.
- Ramon, L et al. (2006). Retrieval, Reuse, Revision, and Retention in Case-Based Reasoning. *The Knowledge Engineering Review*, 20(3), 215-240.
- Reinartz & Roth. (2000). On Quality Measures for Case Base. *In Proceedings of the 5th European Workshop on Case-Based*.
- Rupali, k & Deepali, S. (2018, June 02). APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN HUMAN LIFE. *International Journal of Research*, 6(6).
- Saniya et al . (2011). Fuzzy logic: A “simple” solution for complexities in neurosciences? *Surgical neurology international*.
- Shadia, Y. (2018, April). Applying Case Based Reasoning on Breast Cancer Data: Gaza Strip Case Study. *Masters thesis*.
- Shahina, B et al. (2009, March). Case-based systems in health sciences - a case study in the field of stress management. *WSEAS TRANSACTIONS on SYSTEMS*(ISSN: 1109-2777).
- Shahina, B. (2011, JULY). Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments. *EEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, 41.
- Shelly and Anand. (2011, Apr-May 2). Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2, 188-195.
- Sinan, E et al. (2014, June). Pneumonia: diagnosis and management of community- and hospital-acquired pneumonia in adults. *National Collaborating Centre and NICE project team Guideline*, 1-20.
- Siraj, M. (2019, March 08). *I.J. Education and Management Engineering*, 44-58.
- Sivanandam and Sumathi. (2006). Data Mining: Tasks, Techniques and Applications, in *Introduction to Data Mining and its Applications*.
- Smriti & Nishi. (2014, April). Firewall and Its Policies Management. *International Journal of Computer Science and Mobile Computing*, 3(4).
- Smyth & McKenna. (2001). Competence models and the maintenance problem. 235-249.
- Solomon, G. (2013, JANUARY). A self-learning knowledge based system for diagnosis and treatment of diabetes. *Unpublishe master thesis*.
- Soumen, C et al. (2009). *Data Mining Know It All*. United States: Morgan Kaufmann.

- Speel, P et al. (2001). Conceptual Modelling for Knowledge-Based Systems. *Encyclopedia of Computer Science and Technology*, Marce Dekker Inc., New York, 107-132.
- Sudhir & Kodge. (2013). Census Data Mining and Data Analysis using WEKA. *International Conference in “Emerging Trends in Science, Technology and Management*, (pp. 35-40). Singapore.
- Surabhi & Seema. (2014). Medical Data Mining Life Cycle and its Role in Medical Domain. *International Journal of Computer Science and Information Technologies*, 5 (4).
- Tagel, A. (2013, January). Knowledge based system for pre-medical triage treatment at adama university asella hospital. *Unpublished masters thesis*.
- Tamir, A et al. (2017, September 08). Recommender System in Tourism Using Case based Reasoning Approach. *I.J. Information Engineering and Electronic Business*, 34-43.
- Tesefahun, B. (2012, JUNE). Application of data mining for predicting adult mortality. *Unpublished thesis*.
- Teshome & Hailu. (2017, June 26). Patients’ satisfaction towards radiological service and associated factors in Hawassa University Teaching and referral hospital, Southern Ethiopia. *BMC Health Services Research volume*.
- Teshome.M. (2015, September). Application of casebased recommender system in tourism site selection. *Unpublished mastres thesis*.
- Thangaraj & Vijayalakshmi. (2013, March). Performance Study on Rule-based Classification Techniques across Multiple Database Relations. *International Journal of Applied Information Systems (IJAIS)*, 5, 1-7.
- UNICEF. (2019, November). *UNICEF Data: Monitoring the situation of children and women*. Retrieved July 26, 2020, from UNICEF : <https://data.unicef.org/topic/child-health/pneumonia/>
- Vaishali et al. (2014, March 03). Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis. *INTERNATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH & EMERGING TECHNOLOGIES*, 3.
- Watson & Marir. (1994). Case-Based Reasoning: a Review. *The Knowledge Engineering Review*.
- Wendwesen, E. (2016). Developing a case basd credit approval system using data mining the case of Commercial Bank of Ethiopia. *unpublished master thesis*.

- WHO. (2020, August 12). *Pneumonia*. Retrieved from https://www.who.int/maternal_child_adolescent/news_events/news/2011/pneumonia/en/
- WHO and UNICEF. (2013). The integrated Global Action Plan for the Prevention and Control of Pneumonia and Diarrhoea (GAPPD) 9.
- Wikimedia. (2020, January 13). Retrieved January 15, 2020, from <https://en.wikipedia.org/wiki/Pneumonia>
- Wilke, W. & Bergmann, R. (1998). Techniques and Knowledge Used for Adaptation During Case-Based Problem Solving. *In Proceedings of the 11th International Conference on*.
- Yaswanth & Korrapati. (2016, Nov). Combining Clustering with Classification: A Technique to Improve Classification Accuracy. *International Journal of Computer Science Engineering (IJCSE)*, 5, 336-338.
- Zewudu, M et al. (2020, May 1). Pneumonia among Under-Five Children in Northwest Ethiopia: Prevalence and Predictors—A Community-Based Cross-Sectional Study. *International Journal of Pediatrics*.
- Zhongzhi, S. (2011). *Advanced Artificial Intelligence*. China: World Scientific Publishing Co. Pte. Ltd.

APPENDICES I

APPENDIX I: INTERVIEW QUESTIONS

The main objective of this interview to acquire knowledge about pneumonia diseases from domain expert that will help for the development of a case based reasoning system for diagnosis and treatment of pneumonia. The interviewer records the respondents' response using pen, pencil and paper. I thank you in advance for your willingness and valuable time.

1. What is Pneumonia?
2. What are the different types of pneumonia? Which types of pneumonia are mostly occurring in under five-year children?
3. What are the risk factors of Pneumonia?
4. What are the common signs and symptoms of pneumonia?
5. What are the major pneumonia a diagnosis procedures that you follow and which one is the crucial for your decision making process?
6. How do you identify the major symptoms of pneumonia?
7. Does pneumonia have stages? If it has, what are they and by what measurement they are differentiated?
8. If the compliant is a pneumonia patient, what things are considered by the clinician in order to treat the disease?
9. What are the main decisions that the clinicians make in pneumonia treatment? Or what are the major important actions that you take in to account?
10. Which attribute are the most important in diagnosing the disease that the clinician should focus pneumonia measurement?
11. Do you have standardized guidelines that you use for the diagnosis of pneumonia patients?
12. If your answer for question number 11 is yes, when and in which case you use the guideline?
13. What are the major difficulties and challenges in pneumonia diagnosis? And how can you manage them?

APPENDIX II: USER ACCEPTANCE TEST

This is an evaluation form to be filled by pneumonia diagnosis experts in order to evaluate the applicability of the prototype case based reasoning system for diagnosis and treatment of pneumonia under-five year children. I thank you in advance for your willingness and valuable time.

Description of the parameter values: Performance Value 1=Poor; 2=Fair; 3=Good; 4=Very good; and 5=Excellent

Instruction: Please, tick on the appropriate value for the corresponding parameter of the case based reasoning system for diagnosis and treatment of pneumonia for under-five year children.

No	Evaluation Criteria	Performance Value					
		1	2	3	4	5	Average
1	Is the Prototype system ease to use						
2	Is the Prototype system adequate and clear for decision support?						
3	Relevance of attributes in representing the pneumonia for under-five year children case						
4	Fitness of the final solution to the problem at hand						
5	Relevance of the retrieved case in the diagnosis and treatment of Pneumonia under-five year children						
6	Efficiency of the system in time						
7	Is the prototype system has adequate resource?						
8	Is the prototype system user interface interactive?						
9	Rate the significance of the system in the domain area						
Total Average							

APPENDIX III: PART CLASSIFIER OUTPUTS

=== Run information ===

Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1

Relation: ALL DATA SET weka new org
weka.filters.unsupervised.instance.RemovePercentage-P20.0

Instances: 1291

Attributes: 25

Sex
Age
Cough that produces mucus
Cough
Vomit
Diarrhea
Loss of appetite
Tiredness
Fever
Chills
Fast breathing
Chest pain
Headache
Sweating
Difficulty swallowing
Nausea
Dizziness
Fussiness
Shortness of breathing
Skin rash
Wheezing
Muscle pains
Convulsions
Place
Classification

Test mode: split 80.0% train, remainder test

=== Classifier model (full training set) ===

PART decision list

Convulsions = Yes AND

Headache = No AND

Difficulty swallowing = No: Bacterial-Pneumonia (548.0/3.0)

Muscle pains = Yes AND

Skin rash = Yes: Fungal-Pneumonia (50.0)

Headache = Yes AND

Nausea = No AND

Loss of appetite = Yes AND

Place = CAP: Viral-Pneumonia (103.0)

Chills = No AND

Headache = No AND

Diarrhea = No AND

Loss of appetite = No AND

Difficulty swallowing = Yes: Aspiration-pneumonia (213.9)

Chills = Yes AND

Vomit = No AND

Chest pain = No AND

Place = HAP: Streptococcus-pneumoniae (41.0)

Headache = Yes AND

Tiredness = Yes AND

Difficulty swallowing = No: Mycoplasma-Pneumonia (82.0)

Sweating = Yes AND

Fussiness = No AND

Chest pain = No: Broncho-Pneumonia (17.0)

Chills = No AND
Headache = No AND
Cough = No AND
Loss of appetite = No AND
Chest pain = Yes: Aspiration-pneumonia (24.1)
Dizziness = Yes AND
Fussiness = Yes: Aspiration-pneumonia (11.0/1.0)
Dizziness = Yes AND
Loss of appetite = Yes: Broncho-Pneumonia (10.0)
Chills = No AND
Skin rash = Yes: Aspiration-pneumonia (18.0/1.0)
Chest pain = Yes AND
Loss of appetite = No: VA-Pneumonia (8.0)
Wheezing = Yes AND
Loss of appetite = Yes: Bacterial-Pneumonia (22.0/1.0)
Place = HAP: Bacterial-Pneumonia (15.0/2.0)
Chills = No AND
Fussiness = Yes: Viral-Pneumonia (13.0)
Chills = No AND
Cough that produces mucus = Yes AND
Tiredness = No: Viral-Pneumonia (6.0)
Chills = No AND
Fast breathing = No: Aspiration-pneumonia (23.0)
Wheezing = Yes AND
Nausea = Yes: Aspiration-pneumonia (13.0)
Chills = No AND
Loss of appetite = No AND

Wheezing = No AND

Fever = HGF AND

Age = (0, 1] AND

Tiredness = No AND

Sex = F: Viral-Pneumonia (11.0/1.0)

Chills = Yes AND

Vomit = Yes: Viral-Pneumonia (18.0/1.0)

Chills = Yes AND

Age = (1,3]: Streptococcus-pneumoniae (6.0)

Chills = No AND

Diarrhea = No AND

Cough that produces mucus = No AND

Fever = HGF AND

Headache = No: Viral-Pneumonia (11.0/1.0)

Chills = No AND

Vomit = Yes: Bacterial-Pneumonia (9.0)

Tiredness = No AND

Age = (1,3]: Aspiration-pneumonia (5.0)

Tiredness = Yes AND

Cough that produces mucus = No: VA-Pneumonia (6.0)

Cough that produces mucus = No AND

Fever = LGF: Aspiration-pneumonia (3.0)

Cough that produces mucus = No: Viral-Pneumonia (2.0)

: Bacterial-Pneumonia (2.0)

Number of Rules: 28

Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	254	98.4496 %
Incorrectly Classified Instances	4	1.5504 %
Kappa statistic	0.9781	
Mean absolute error	0.0056	
Root mean squared error	0.0625	
Relative absolute error	3.1549 %	
Root relative squared error	20.9916 %	
Coverage of cases (0.95 level)	98.4496 %	
Mean rel. region size (0.95 level)	13.0329 %	
Total Number of Instances	258	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.991	0.014	0.983	0.991	0.987	0.989	Bacterial-Pneumonia
0.975	0	1	0.975	0.987	0.988	Viral-Pneumonia
1	0.004	0.8	1	0.889	0.998	Broncho-Pneumonia
1	0	1	1	1		Mycoplasma-Pneumonia
1	0	1	1	1		VA-Pneumonia
0.968	0	1	0.968	0.984	0.988	Aspiration-pneumonia
1	0.004	0.909	1	0.952	0.998	Fungal-Pneumonia
1	0	1	1	1		Streptococcus-pneumoniae
Weighted Avg.	0.984	0.007	0.986	0.984	0.985	0.99

=== Confusion Matrix ===

a b c d e f g h <-- classified as

116 0 1 0 0 0 0 0 | a = Bacterial-Pneumonia

0 39 0 0 0 0 1 0 | b = Viral-Pneumonia

0 0 4 0 0 0 0 0 | c = Broncho-Pneumonia

0 0 0 9 0 0 0 0 | d = Mycoplasma-Pneumonia

0 0 0 0 5 0 0 0 | e = VA-Pneumonia

2 0 0 0 0 61 0 0 | f = Aspiration-pneumonia

0 0 0 0 0 0 10 0 | g = Fungal-Pneumonia

0 0 0 0 0 0 0 10 | h = Streptococcus-pneumoniae