
# SCHOOL OF GRADUATE STUDIES

# DEVELOPING CLASSIFICATION MODEL WITH KNOWLEDG BASE SYSTEM FOR DIAGNOSIS AND TREATMENT RECOMMENDATION OF HOSPITAL ACQUIRED PNEMONIA

**MSc. THESIS**

**WONDIMU KIBATU GIRMA**

**APRIL, 2024**

**WOLKITE, ETHIOPIA**

# WOLKITE UNIVERSITY
# SCHOOL OF GRADUATE STUDIES

**Developing Classification Model with Knowledge Base System for Diagnosis and Treatment Recommendation of Hospital Acquired Pneumonia (HAP)**

**A Msc Thesis Submitted to School of Graduate Studies, in Partial Fulfillment of Requirement for the Degree Masters of Science in Computer Science and Engineering (Specialization: Computer Science)**

**Wondimu Kibatu Girma**

**Major Advisor: Kindie Biredagn (Ph.D.)**

**Co_Advisor: Abunu Tesfaw (M. Sc)**

**April, 2024**

**Wolkite, Ethiopia**

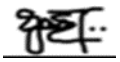
**SCHOOL OF GRADUATE STUDIES**

**WOLKITE UNIVERSITY**

**ADVISORS' APPROVAL SHEET**

---

This is to certify that the thesis entitled "**Developing Classification Model with Knowledge Base System for Diagnosis and Treatment Recommendation of Hospital Acquired Pneumonia**" submitted in partial fulfilment of the requirements for the degree of **Master's** with specialization in **Computer Science and Engineering**, the Graduate Program of the **Department of** Software Engineering, and has been carried out by **Wondimu Kibatu Girma** Id **CIGR/013/14** Under our supervision. Therefore, we recommend that the student has fulfilled the requirements and hence hereby can submit the thesis to the department.

| | | |
|---|---|---|
| **Kindie Biredagn (Ph.D.)** |  | April-12-2024 |
| Name of major advisor | Signature | Date |
| **Mr. Abunu Tesfaw (M.Sc.)** |  | April-14-2024 |
| Name of Co-advisor | Signature | Date |

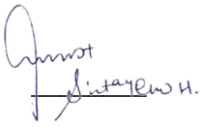
**SCOOL OF GRADUATE STUDIES**

**WOLKITE UNIVERSITY**

**EXAMINERS' APPROVAL SHEET**

---

As members of the Board of Examiners of the Master's degree open defense, we certify that we have read and evaluated the thesis entitled '' "**Developing Classification Model with Knowledge Base System for Diagnosis and Treatment Recommendation of Hospital Acquired Pneumonia**" prepared by Wondimu Kibatu Girma and Examined the candidate. And recommend that it be accepted as fulfilling the thesis requirement for the degree of Master of Science (M.Sc.) in Computer science and engineering.

| | | | |
|---|---|---|---|
| Sintayehu H. (PhD) | External examiner |  | 27-May-2024 |
| | Internal examiner | | |
| | Chairperson | | |

Final approval and acceptance of the thesis is contingent upon the submission of the final copy of the thesis to the SGS through the DGC/SGC of the candidate's department/School.

## DECLARATION

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

---

**Wondimu Kibatu Girma**

**April, 2024**

The thesis has been submitted for examination with our approval as university advisor.

---

**Dr. Kindie Biredagn**

**April, 2024**

## DEDICATION

This thesis is dedicated to my family and friends who supported me on my success.

# BIOGRAPHY

The author, Mr. Wondimu Kibatu, was born on November 07, 1987 E.C, at Shamene Kebele Guraghe Zone, Central Ethiopia Region, Ethiopia, from his father, Ato Kibatu Girma, and mother Woizero Zuriyash Nida. He attended his primary education at worit primary school, his secondary education at Agena secondary and preparatory school. Then he joined Wolkite University department of Information System in 2006 E.C and obtained his BSc degree in Information System in June, 2009 E.C. He was employed in Wolkite University in 2010 E.C. Then, in November, 2014, he joined the department of Software engineering, in college of computing and informatics at Wolkite University to pursue his Master of Science degree in computer science and Engineering.

# ACKNOWLEDGEMENT

Above all, I want to thank Almighty God from the bottom of my heart for supporting me throughout this thesis, from the beginning to the end, and for doing it in such a lovely way.

I desire to express my gratitude to Dr. Kindie Biredagn, my research advisor, for his informative and helpful advice. He gave me a lot of inspiration to work on this study. His readiness to inspire me made a significant contribution to the study. I have no words to express how much I appreciate your understanding and committed extremely knowledgeable support. In addition, I would like to take this opportunity to thank my co-advisor Mr. Abunu Tesfaw for his deep guidance and comments in all of my work.

Additionally, I owe a special thank you to the staffs of Werabe Referral Hospital and Wolkite University Specialized Teaching Hospital for their assistance with the user acceptance assessment, system performance testing, and knowledge acquisition process, Redi Mohammed, Shamil Kenzu, for their support, interest, and dedication to knowledge sharing.

Additionally, it gives me a lot of joy to thank wolkite university management members for their support and encouragement as I completed this thesis work. Moreover, I like to highly acknowledge the affection and continued encouragement of Mr. Tesfaye minuta (Asst.Prof) Wolkite University registrar directorate director for his dedication and kind support.

Lastly, I would want to express my deep gratitude to all of my friends for their invaluable support throughout my entire education journey.

Wondimu Kibatu

April, 2021

# Table of Contents

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

## ABBREVIATIONS DESCRIPTION

| | |
|---|---|
| **AoR** | Adjusted Odds Ratio |
| **CAP** | Community Acquired Pneumonia |
| **CBR** | Cased Based Reasoning |
| **CLR** | Community Library Programme |
| **CRISP** | Cross Industry Standard Process |
| **CRISP** | Cross Industry Standard Process |
| **CxR** | Chest X-Ray |
| **(CLI)** | Command line interface or |
| **DM** | Data Mining |
| **DoR** | Diagnostic Odds Ratio |
| **DSRPM** | Design Science Research Process Model |
| **EM** | Expectation Maximization |
| **EK** | Expert Knowledge |
| **(GUI)** | Graphical User Interface |
| **HAP** | Hospital Acquired Pneumonia |
| **HR** | Heart Rate |
| **ICU** | Intense Care Unit |
| **JDK** | Java Development Kit |
| **KBS** | Knowledge Based System |
| **KDD** | Knowledge Discovery in Data Base |
| **KR** | knowledge representation |
| **LR** | Logistic Representation |
| **MRSA** | Methicillin Resistant Staphylococcus Aureus |
| **NLR** | Negative Likelihood Ratio |
| **Osat** | Oxygen Saturation |

| | |
|---|---|
| **PEM** | Protein-Energy Malnutrition |
| **PLR** | Positive Likelihood Ratio |
| **RBR** | Rule Based Reasoning |
| **RIPPER** | Repeated Incremental Pruning to Reduce Error Reduction |
| **ROI** | Regions of Interest |
| **SE** | Sensitivity |
| **SoB** | Shortness of Breath |
| **SP** | Specificity |
| **SVM** | Support Vector Machine |
| **TB** | Tuberculosis |
| **US** | United State |
| **VF** | Vocal Fremitus |
| **VCT** | Volunteer Counseling Test |
| **WEKA** | Waikato environmental for knowledge Analysis |

# ABSTRACT

Pneumonia is an illness, usually caused by infection, in which the lungs become inflamed and congested, reducing oxygen exchange and leading to cough and breathlessness. It affects individuals of all ages but occurs most frequently in children and elderly. Pneumonia has different categories. Hospital Acquired Pneumonia (HAP) is the first in mortality and morbidity. And lack of health facilities, lack of sufficient professional in hospitals and complexity of the diagnosis process are problems exposed. KBS has great role in the health care sector so this study aims to combine data mining results with expert knowledge, and establish a Knowledge Base System for the diagnosis and treatment recommendation of HAP. Design science research methodology, with a hybrid data mining process model was employed. The researcher gathered a dataset of 3244 cases of Hospital Acquired Pneumonia (HAP) from Werabe Referral Hospital.

The random forest, J48, JRip, and PART algorithms were used in 4 tests with two distinct scenarios by the researcher in order to create the classifier model. PART classifier algorithm conducted on selected attributes with percentage split test option with an accuracy of 99.3% was achieved. To model the gained knowledge decision tree modeling technique and to represent the gained knowledge the rule-based knowledge representation technique was used. Semi structured interview technique is chosen for acquiring knowledge from expert. Then it is modeled by using the decision tree modeling techniques and represented in the production rule. The two extracted knowledge was combined and checked for rule redundancy to develop the knowledge-based system.

Finally, to develop the KBS the researcher used SWI prolog and Net Beans for making user interface. To evaluate performance of the developed system, the researcher has used system performance testing and user acceptance evaluation. And Achieves 90.3% accuracy for system performance. And 91.3% of accuracy for user acceptance testing. The result show that the developed system achieves good performance and meets the objectives of the study and it could give proper treatment. This deduces that the developed system could help in identifying the severity level and in diagnosis and treatment recommendation of Hospital Acquired Pneumonia.


***Keywords:*** Pneumonia, Data Mining, HAP, Knowledge-Based System

# CHAPTER ONE
# INTRODUCTION

## 1.1 Background

Hippocrates (460–370 BC) was the first to describe pneumonia. Rokitansky and Laennec provided the first descriptions of its clinical and pathological characteristics in 1819, which was 22 centuries later [1]. Alveoli, which are tiny air sacs in the lungs, fill with air when a healthy individual breathes. When someone has pneumonia, their alveoli are packed with pus and fluid, which makes breathing challenging and lowers oxygen intake [2]. The leading infectious cause of death in children worldwide is pneumonia. In 2019, pneumonia claimed the lives of 740180 children under the age of 5, accounting for 14% of all pediatric deaths but 22% of all deaths [3]. A complicated series of events lead to the illness known as pneumonia, which starts with the first contact with a pathogenic microbe and ends with the invasion of the lower respiratory tract.

This infection can be contracted outside of a medical setting as well. It can also be spread through inhaled or aspirated bacteria. Additionally, it is a major global health issue that contributes significantly to morbidity and mortality. It causes nearly 14,000 hospital readmissions, 50,000 deaths, and 1.1 million hospital admissions annually in the US alone[4]. Bacterial, viral, or fungal infections can cause pneumonia (but most commonly bacterial). In order to properly manage and direct appropriate treatment for pneumonia, it is crucial to comprehend the involvement of the many pathogens in the microbiological etiology of the disease.

In the medical field, data mining techniques are typically integrated with rule-based or case-based reasoning systems [5].In this study, knowledge-based system (KBS) with a classification model and a rule-based reasoning system are developed for hospital acquired pneumonia. The dataset will be gathered from the pediatric ward of Werabe Hospital, and will be preprocessed using data mining tools.

Classification data mining technique is a method that identifies shared characteristics among the attributes of database items and divides them into several classes that reflect the training data set as a learning model [5] [6] Because continually classify, categorize, and rate the nature of the collections of items or erroneous information around us in order to comprehend and communicate about the world, classification is one of the most frequent data mining activities in our daily lives.

## 1.2 Background of the study

Hospital-acquired (nonsocial) pneumonia (HAP) is a pneumonia that occurs 48 hours or more after admission, Alveoli, which are tiny air sacs in the lungs; fill with air when a healthy individual breathes [3]. When someone has pneumonia, their alveoli are packed with pus and fluid, which makes breathing challenging and lowers oxygen intake. The leading infectious cause of death in children worldwide is pneumonia. Affecting 0.5 to 1.7% of hospitalized patients HAP is the most lethal of all hospital nosocomial infections and the most common cause of mortality. A HAP account for 50% of all antibiotics administered in the hospital setting and has significant impact on health care costs. HAP is a dynamic illness with many etiologies, a fluctuating natural history, and a host of risk factors[5].

The severity of the illness should always be assessed in patients who have been diagnosed with hospital acquired pneumonia [3]And it has three severity levels namely mild, moderate and severe. Predictive patient analysis, used in the majority of currently accessible cases, assesses illness severity and aids in the diagnosis of pneumonia [3]. It also helps to determine the necessity for an etiological study, the type of antibiotics to be used, and how to administer them. The likelihood of oral medication usage, comorbidities, psychosocial issues, and socioeconomic status are only a few of the variables that affect how severe the condition is. However, the most important variable that must be taken into account when making a decision is the individual. Etiological studies are not required for patients getting outpatient care for non-severe pneumonia, according to [5].

However, the current ICT intervention, particularly a combined knowledge-based system, a family of artificial intelligent systems, plays a crucial role in resolving issues related to mortality and morbidity as well as any other medical issue by assisting with data analysis, diagnosis, and treatment. This drives the need for this study, which uses data mining software integrated with a combined knowledge-based system to analyze patient histories and clinical datasets from Werabe Comprehensive Specialized Hospital for the diagnosis and treatment of pediatric community acquired pneumonia. Childhood pneumonia is one of the leading causes of morbidity and mortality among children under the age of five worldwide.

South-central Ethiopia contains the town of Werabe. This town is located in the Silte Zone of the Southern Nations, Nationalities, and Peoples Region, according to official sources. And has its own health centers. Among these werabe comprehensive hospital is one of them that provide different health services with specialized experts. Among those services treatment of pneumonia is one of them and at regional level this hospital is ranked in provide such kind of diseases but there are different problems to get service or treatment and diagnosis. Such as lack of dormitories financial problem etc.

## 1.3 Motivation of the Research

Pneumonia is an illness, usually caused by infection, in which the lungs become inflamed and congested, reducing oxygen exchange and leading to cough and breathlessness. it affects individuals of all ages but occurs most frequently in children and elderly. And it is the most common cause of mortality and morbidity worldwide. In developing nations such as Ethiopia it is the most common and challenging in hospital admission, diagnosis and treatment process. Diagnosis and treatment of Hospital Acquired Pneumonia (HAP) requires navigating a complex of symptoms, agents of infection, and patient-specific features and the diagnosis process is highly dependent on experienced professionals. The misdiagnosis by physicians, the lack of qualified medical professionals, and the delay in diagnosis and treatment are the causes that contributed to the higher death rate. Based on those reasons and the stated problem listed below the researcher initiated to implement a new strategy for medical facilities that can accurately

diagnose and provide prompt, appropriate treatment, and lower the death rate and morbidity rate.

## 1.4 Statement of the Problem

Pneumonia is the most common causes for body blood pressure to drop to dangerous levels called septic shock causing 50% of all episodes in the world. Given the poor disease surveillance and overall inadequate health systems in poorer nations, several researchers claim that hospital acquired pneumonia is the top cause of death in children under five years old [7]

The underlining research problem that necessitated is existence of high death rate of pneumonia at national level.it is reached about 59931 deaths accounting for 10.63% of total death. Pneumonia contributes for the occurrence of other illness such as chronic obstructive pulmonary diseases, heart diseases, kidney diseases, diabetes and HIV Aids [8]. Although pneumonia in children under the age of five can be detected and diagnosed earlier, it still kills over 2400 people every day and kills 100 children every hour [8].

Following India, Nigeria, Pakistan, and the Democratic Republic of the Congo, which together account for 49% of all pneumonia-related deaths worldwide, Ethiopia is one of the top five nations in the world for the number of children under the age of five who die [7], [8]There are associated risk factors for pneumonia such as socio demographic status (educational status, occupation, monthly income, and family size), living condition factors example housing condition, toilet facility, water facility and cooking area. Thus, given the primary cause of baby deaths is hospital-acquired pneumonia, which should be treated before it progresses to a life-threatening stage and with the proper medication. To identify and to make correct decision, it requires properly organized knowledge during consultation, as well as diagnosis and at the time of treatment [7].

In rural areas of Ethiopia Peoples are enforced to go a much longer way for accessing health related services because of lack of health facilities [8]. At the time of ward patients are enforced to left the hospital before properly and regularly completing their required medication and

required treatment because of lack of dormitories [9]. Most of the time, lack of sufficient professional in hospitals and other health sectors is the main problem that makes patients to keep and wait the service in shifts for a long period of time one after another to complete the treatment and it is boring for the patient and makes them hopeless on their health-related issue [10].

The use of data mining and machine learning techniques in disease diagnosis and treatment has become increasingly important due to several factors: With the advent of digitization and electronic health records, an enormous volume of medical data is being generated. This includes patient histories, diagnostic test results, and other relevant information. Data mining and machine learning allow us to extract meaningful patterns and insights from this vast amount of data. By analyzing large medical datasets, these techniques can identify subtle patterns and correlations that may be missed by human observation. Early detection of diseases is crucial for timely intervention and better outcomes.

Complexity of the diagnosis process such as symptom similarity of pneumonia with other diseases, lately Detection and Prevention and the scarcity of qualified medical personnel is one of the main obstacles for pneumonia diseases [6] [11]. According to, there is a 4.3 million-worker deficit in healthcare institutions worldwide. According to [6]. The report finds that in Ethiopia, there are 0.84 healthcare personnel for every 1000 people. Despite having the most healthcare workers per capita in Sub-Saharan Africa, Ethiopia falls short of the World Health Organization (WHO) standard of 2.28 healthcare workers for every 1000 people [6].

According to werabe comprehensive hospital reports currently, Hospital acquired pneumonia was the first of any other diseases. Data mining techniques with machine learning algorithm should be applied for the diagnosis and treatment for hospital acquired pneumonia that can contribute to eliminating the differences in knowledge among health professionals [2] and to improve the quality-of-service delivery in the health care institution.

"Severe pneumonia or extremely serious disease" was the classification given to kids who displayed any overt danger indications [10]. Based on this knowledge, the World Health

Organization (WHO, 2016) reviewed the evidence in order to create a more straightforward strategy that would enhance the proportion of kids receiving the right treatment for pneumonia.

Currently, the healthcare system's top priority is patient pleasure, especially in emerging nations. Additionally, patient satisfaction has developed into a recognized outcome indicator, a tool to assess healthcare quality, and a source of information for developing strategies for affordable, sustainable, Patient treatment that is acceptable and affordable [10]. The patient dataset could be used to apply a wealth of specialized knowledge to generate effective, dependable, and high-quality services across all health sectors.

After analyzing the above stated papers and by investigating the current diagnosis and treatment process of hospital acquired pneumonia (HAP) in the selected organization the overall diagnosis and treatment process is complex, such as **similar symptoms** can overlap with other respiratory conditions like common cold or flu. **Clinical assessment** physicians rely on physical exams and patient history, but these may not always provide a definitive diagnosis and treatment. **Severity assessment** determining the severity of the diseases guides treatment decisions. Some patients may need hospitalization, while others can be managed at home. **Duration of Treatment** balancing adequate treatment duration with avoiding unnecessary antibiotic exposure is important task. Based on this stated dilemma the researcher proposed classification model with knowledge-based system for diagnosis and treatment recommendation of Hospital Acquired Pneumonia (HAP) for providing the right service at the right time to the right person.

At the end, the study answers the following question

**RQ 1:** What are the determinants attributes for the diagnosis and treatment of hospital acquired pneumonia (HAP) diseases?

**RQ 2:** Which classification algorithm is the best to develop prediction model for hospital acquired pneumonia (HAP) diseases?

**RQ 3:** What knowledge is obtained from experts and data mining for developing Knowledge base system?

## 1.5 Objectives of the Study

### 1.5.1 General Objective

The general objectives of this study are to develop classification predictive model with knowledge-based systems for diagnosis and treatment of hospital acquired type of pneumonia (HAP) the case of Werabe comprehensive hospital.

### 1.5.2 Specific Objective

- To Acquire knowledge from expert.
- To preprocess the dataset using data preprocessing tools and techniques.
- To perform data transformation.
- To Select the determinant attributes/features for the diagnosis of the selected diseases.
- To apply classification techniques to build the model using classification algorithms.
- To evaluate the performance of the algorithms using different evaluation metrics.
- To acquire knowledge from domain expert.
- To assess the severity levels of Hospital Acquired Pneumonia (HAP) diseases.
- To develop a knowledge base system prototype by using rule-based reasoning.
- To provide a conclusion and recommendation based on the study result.

## 1.6 Scope and Limitation of the Study

The scope of this research is to develop Classification model with knowledge base system for diagnosis and treatment recommendation of Hospital acquired pneumonia (HAP) to support the clinical decision-making process of werabe referral Hospital.

There are different types of pneumonia such as ventilator associated pneumonia and community acquired pneumonia but the study is mainly limited on hospital acquired type of pneumonia which accounts about 20% of mortality which is approximately twice that of community acquired type of pneumonia. Because of shortage of resource and time and I will collect pneumonia diseases data from werabe comprehensive hospital starting from 2007 e.c and the total number of data and their features briefly discussed on chapter three.

## 1.7 Significance of the Study

A knowledge-based system (KBS) is a type of computer system that examines information from several sources, including knowledge, data, and other information, in order to produce new knowledge. It uses AI concepts to solve problems, offer guidance and consultation, which may be useful for assisting with human learning and making decisions. Knowledge-based systems increase productivity and improve the ability to solve problems in a flexible way when knowledge is used properly. Additionally, it can be used to record knowledge for later use, which enhances the quality of the problem-solving process. Therefore, creating and deploying a knowledge base system will help to diagnose and treat hospital acquired pneumonia (HAP) effectively and will bring both direct and indirect benefits.

The specialists involved in diagnosing pneumonia patients in hospitals and healthcare facilities will directly benefit from this research's findings. When diagnosing and treating pneumonia patients, the system aids professionals in regulating and controlling the consistency of the infections. Additionally, it would be beneficial for health researchers or experts to raise decision-makers' awareness of the issue of Hospital Acquired Pneumonia (HAP). These systems also make effective use of expert knowledge sources and enable the documentation of one or more of them. Patients who are diagnosed with Hospital Acquired Pneumonia (HAP) and their families, teachers, friends, and leaders in their fields who want to raise awareness of the disease's complexity, early detection methods, and long-term treatment options are the indirect benefactors of this research's output. It benefits healthcare professionals working in all levels of the healthcare system since the combined knowledge base system (Knowledge from data mining and Expert) is already set up to give the right therapy to the right patient at the right time. Reduce the amount of time spent on diagnosing might be managed at home and improve local medical care to increase the timeliness of service. Consequently, the suggested approach aids in reducing illness transmission throughout the population and offers effective diagnosis and treatment.

After the research is completed

- Implementing a knowledge-based system could reduce the workload for human experts and shorten hospital patient wait times.

- The knowledge-based system could help human specialists by giving them the information they need to make decisions at the appropriate moment.
- It offers a novel method for the identification and treatment of Hospital Acquired Pneumonia (HAP) diseases.
- The knowledge base developed would contribute for increasing awareness in the issue of diagnosis and working experience of professional.

## 1.8 Methodology of the Study

### 1.8.1 Data collection Method

The researcher has employed both primary and secondary sources of knowledge to obtain the needed knowledge. Domain specialists were contacted for primary data utilizing both structured and Unstructured interviewing techniques. In contrast, secondary data was gathered from the pneumonia patient Card at Werabe comprehensive specialized Hospital.

### 1.8.2 Data mining process method

Because it combines elements of the Cross Industry-Standard Process (CRISP-DM) and knowledge Discovery from the database (KDD), the hybrid data mining process model was chosen. It was created Utilizing the CRISP-DM methodology.

The processes are described in a Broader, more research-focused Manner and data mining activities are included in place of the modeling phase. Thus, it entails six steps: Domain comprehension, data comprehension, data preparation, Data mining, assessment, and knowledge discovery.

### 1.8.3 Implementation tools and techniques

After the dataset is prepared in a way that is appropriate for data mining techniques, it is preprocessed in this Research work, handling missing values using Ms Excel before being analyzed Using the WEKA (Waikato Environment for Knowledge Analysis) Version 3.8.2 for Knowledge Analysis. The researcher utilized SWI-PROLOG version 7.6.4 Programming

language to construct a prototype knowledge-based system after extracting the hidden information from the pre-processed dataset and evaluating classifier performance.

### 1.8.4 Evaluation method

By comparing the models' accuracy using confusion matrices such as accuracy, true positive rate, True negative rate, precision, and F-measure, the researcher in this study assessed the data mining model and evaluate the developed prototype by the end user.

## 1.9 Organization of the study

This research is organized in to six chapters. The **first chapter** briefly discussed the introduction about background, background of the study, statement of the problem, the objective of the study, the scope and limitation and significance of the study. The **second chapter** is all about reviewing the literature. Regarding data mining and knowledge base system technologies, techniques, and algorithms, as well as the DM process and its various duties as applied to the health care industry. On how the research was carried out, including business comprehension, data interpretation and selection, preprocessing of the data, explanation of the classification method used, and detailed discussion on how to evaluate the model covered in **chapter three**. The topics of business comprehension, data comprehension, data preparation and data preprocessing for experimentation were covered in **chapter four.** The experiments in **Chapter Five** compare each algorithm, choose the best algorithm based on the rules chosen, and test the algorithm's acceptability with users. Implementation and discussion of result, a brief overview of knowledge-based system building, system evaluation, and findings discussion are discussed in **Chapter Six.** The research project's conclusions and recommendations are presented in **chapter Seven.**

# CHAPTER TWO
# LITERATURE REVIEW

The review of the literature on pneumonia, with a focus on hospital acquired pneumonia is the topic of this chapter. Additionally, it elaborates the notions of knowledge-based systems, the KBS's architecture, methods for knowledge representation, and tasks of tools for developing knowledge-based systems, data mining, and evaluation measures as well as developing classification model. The discussion of a review of related publications, which is used to identify research gaps, concludes this chapter.

## 2.1 Background of Pneumonia

Acute lung inflammation brought on by an infection is known as pneumonia. Chest X-rays and clinical data are typically used to make the initial diagnosis. If the infection is bacterial, mycobacterial, viral, fungal, or parasitic, whether it is acquired in the community or a hospital, whether the patient is receiving mechanical ventilation, and whether the patient is immune competent or immune compromised, different causes, symptoms, treatments, preventive measures, and prognoses apply [12].

### 2.1.1 Ventilator-Associated Pneumonia (VAP)

After endotracheal intubation, ventilator-associated pneumonia (VAP) begins to manifest at least 48 hours later. Gram-negative bacilli and Staphylococcus aureus are the most prevalent pathogens; antibiotic-resistant organisms are a serious issue. Pneumonia often shows up in ventilated patients as a fever, rise in white blood cell count, decline in oxygenation, and an increase in tracheal secretions, some of which may be purulent. A positive blood culture for the same pathogen detected in respiratory secretions or bronchoscopic sample of the lower respiratory tract with quantitative Gram stain and cultures may be used to confirm a diagnosis that is suspected based on clinical presentation and a chest x-ray. Antibiotics are used as a treatment. The prognosis overall is poor, partially because of comorbidities [13].

### 2.1.2 Community Acquired pneumonia (CAP)

Pneumonia acquired outside of a hospital is referred to as community-acquired pneumonia. Streptococcus pneumonia, Hemophilus influenza, atypical bacteria (such as Chlamydia pneumonia, Mycoplasma pneumonia, and Legionella species), and viruses are the most often found pathogens. Fever, cough, sputum production, leuritic chest discomfort, dyspnea, tachypnea, and tachycardia are symptoms and indicators. The clinical presentation and chest x-ray are used to make the diagnosis. Antibiotics are chosen based on empirical data for treatment. Although much pneumonia, particularly those brought on by Legionella, Staphylococcus aureus, or influenza virus, is dangerous or even fatal in older, sicker patients, the prognosis is excellent for relatively young or healthy people [14].

### 2.1.3 Hospital-Acquired Pneumonia (HAP)

Hospital-acquired pneumonia (HAP) develops at least 48 hours after hospital admission. The most common pathogens are gram-negative bacilli and Staphylococcus aureus; antibiotic-resistant organisms are an important concern. Symptoms and signs include malaise, fever, chills, rigor, cough, dyspnea, and chest pain. Diagnosis is suspected on the basis of clinical presentation and chest imaging and is confirmed by blood culture or bronchoscopic sampling of the lower respiratory tract. Treatment is with antibiotics. Overall prognosis is poor, due in part to comorbidities [15].

#### 2.1.3.1 Diagnosis of Hospital Acquired Pneumonia

The prevalence of drug-resistant bacteria, such as MRSA, glucose non-fermenters, such as Pseudomonas aeruginosa, and Enterobacteria, such as Escherichia coli, Klebsiella, and Enterobacter, is a specific issue in hospital acquired pneumonia. Identification of the causing bacterium is essential for efficient treatment of patients with hospital-acquired pneumonia (HAP). In real practice, however, a large number of patients must be treated even while the exact reason is unknown because the necessary specimens cannot be collected or the diagnostic procedures themselves have limits [14].

Hospital-acquired pneumonia (HAP) is thus if two of the following conditions are met in addition to the presence of aberrant and deteriorating shadows on chest radiography, the condition is diagnosed.

- ✓ Fever ≥38 ℃
- ✓ Cough with sputum
- ✓ Dyspnea
- ✓ Pleuritic chest pain
- ✓ Tachypnea
- ✓ Dullness to Percussion
- ✓ Increased Vocal Fremitus
- ✓ Abnormal white blood cell count (increased or decreased)
- ✓ Purulent discharge.

These diagnostic criteria were created with the intention of minimizing false negatives, even if it meant accepting a small proportion of false positives, in light of the observation that treatment delays are associated with elevated death rates. Although symptoms (i)–(iii) are present, bronchitis or trachea bronchitis is still suspected in situations where there are no aberrant shadows on chest radiography. These criteria were created with the intention of identifying hospital-acquired pneumonia as soon as symptoms appear, and the accuracy of the diagnosis must be assessed 2-3 days after the commencement of therapy based on trends in following clinical symptoms and the outcomes of laboratory testing [13] [15].

#### 2.1.3.2 Treatment of hospital acquired pneumonia

Although delaying the start of antibiotic therapy is associated with a higher risk of death, recent studies argue that antibiotics may not be immediately required in every patient with suspected HAP [56]. Two different strategies clinical and bacteriologic can be used. In the clinical strategy, antibiotics are started in patients with a new pulmonary infiltrate concerning for Hospital Acquired Pneumonia (HAP) based on the severity level of the diseases such as Piperacillin /tazobactam 4.5 g IV 8 hourly plus Gentamicin5 mg/kg/day IV, Cephazolin 1 g IV

8 hourly plus Gentamicin* 5 mg/kg/day IV and amoxycillin + clavulanic acid 875/125 mg (1 tablet) orally 12 hourly.

## 2.2 Overview of Data Mining

The process of obtaining or mining data from massive amounts of data is known as data mining. Knowledge mining from data or "Knowledge mining" is better names for the practice known as data mining. Because of advancements in data gathering and storage technologies, businesses can now amass massive volumes of data for less money. The main objective of data mining, a general activity, is to take advantage of this grip on information in order to extract relevant and usable data. Numerous sectors, including business, sports, marketing, astronomy, medicine, and many more, have successfully used data mining techniques.

The popularity of knowledge discovery from vast amounts of data is a result of it. Individual research projects, clinical procedures, community monitoring, and reports from various medical settings, such as test results and patient records, create enormous volumes of medical data every day. These data need to be properly categorized and evaluated in order to be turned into knowledge that can be used to make wise decisions. New computational strategies are needed to handle these enormous data repositories and uncover complex patterns in data that are challenging to analyze with conventional statistical tools [16].

Because there is a need for an effective analytical process for finding undiscovered and important information in health data, data mining is growing in popularity in the healthcare industry. Data mining has several advantages in the health sector, including the capacity to spot health insurance fraud, make affordable medical solutions available to consumers, find the root causes of disorders, and pinpoint treatment options. Information that includes specifics on hospital patients, health insurance claims, treatment costs, etc. In order to analyze and extract crucial information from this complicated data, a potent tool must be created. The analysis of health data enhances patient management activities, which in turn improves healthcare.

Currently, the health care sector produces a lot of complicated data on patients, hospital resources, illness diagnosis, electronic patient records, medical equipment, pharmaceuticals, and human resources. The only health station that provides services is the hospital. Larger volumes of data must be processed and evaluated in order to support cost-saving measures and decision-making that may be accomplished using a variety of ways. Knowledge discovery from Data (KDD) is another name for data mining. Data mining is a technique used to extract valuable information from sizable databases or data warehouses. Data mining is becoming popular in the healthcare industry nowadays since it is a crucial component of operational analytical approach for finding useful and undiscovered information in health data [17].

### 2.2.1 Knowledge discovery in database

In 1996, Fayyad introduced the initial KDD technique, which involves multiple phases that can be completed iteratively. KDD is defined as a complex method for discovering genuine, novel, potentially valuable, and ultimately understandable patterns within data. This popular data mining method includes data selection, preprocessing, transformation, and mining, along with the precise interpretation of findings from observable outcomes. Knowledge discovery in databases is a well-defined process comprising several unique steps [18].

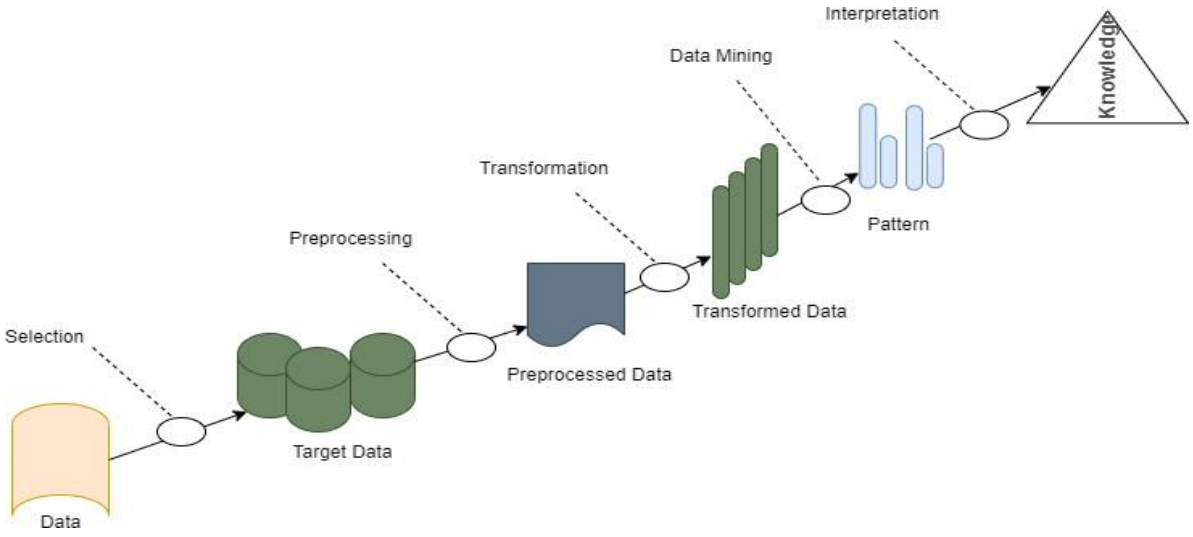



Figure 2.1: Knowledge base system architecture adopted from [35]

Steps of knowledge discovery in the database adopted from [20].

**Selection**

The first step in KDD is data selection where it is the process in which data from the data gathering are selected and made relevant for the study. At this step, a target dataset is being created with a narrow emphasis on the variables that will be used in the discovery process to help address the issue.

**Preprocessing**

To get a consistent dataset and improve the chances of effective data mining, it involves data cleaning and preprocessing operations. Making sure that data records are full, removing or correcting for noise, filling in missing information, etc. are some of the procedures here. It includes data cleaning, such as handling missing quantity handling and removing noise or outliers. In this situation, it may employ a Data Mining method or sophisticated statistical approaches. For instance, the goal of the Data Mining supervised method could change when it is suspected that a certain attribute is unreliable or has a large number of missing data.

**Transformation**

The process of changing data into the right form needed by mining procedures is known as data transformation. The two steps of data transformation are as follows:

I. Data Mapping: Assigning elements from source base to destination to capture transformations.

II. Code generation: Creation of the actual transformation program.

In this phase splitting the data attribute range into intervals that each include roughly the same number of samples or by scaling the attribute data to fall inside a given range, data are consolidated into forms suitable for mining to minimize data size. As a result, in order to facilitate data mining, attribute values are altered to a new set of replacement values.

**Data mining**

Strategies used to extract potentially relevant patterns. It turns task-relevant data into patterns and uses classification to determine the model's purpose. A crucial step in which intelligence techniques are used to uncover hidden patterns in the data. In this stage, the primary issue must

be analyzed for patterns of data interest based on the goals of the company and the needs of data mining. To build predictive or descriptive models, several data mining methods and approaches are employed to look for information or intriguing patterns.

**Interpretation**

At KDD, the mined patterns and connections are interpreted at this post-processing stage. KDD is an iterative procedure because if the pattern assessed is not beneficial, the process may restart from any of the previous phases. This section consists of turning the patterns into knowledge by eliminating useless or unnecessary patterns and putting the beneficial patterns into intelligible (human understandable) language.

### 2.2.2 CRISP-DM process for Data Mining (CRISP-DM)

Is a common process model that may be applied to data mining to explore databases for trends, correlations, and patterns. The standard outlines six distinct steps that must be completed once or more times. Is one of the most extensively used data mining approaches for knowledge discovery.

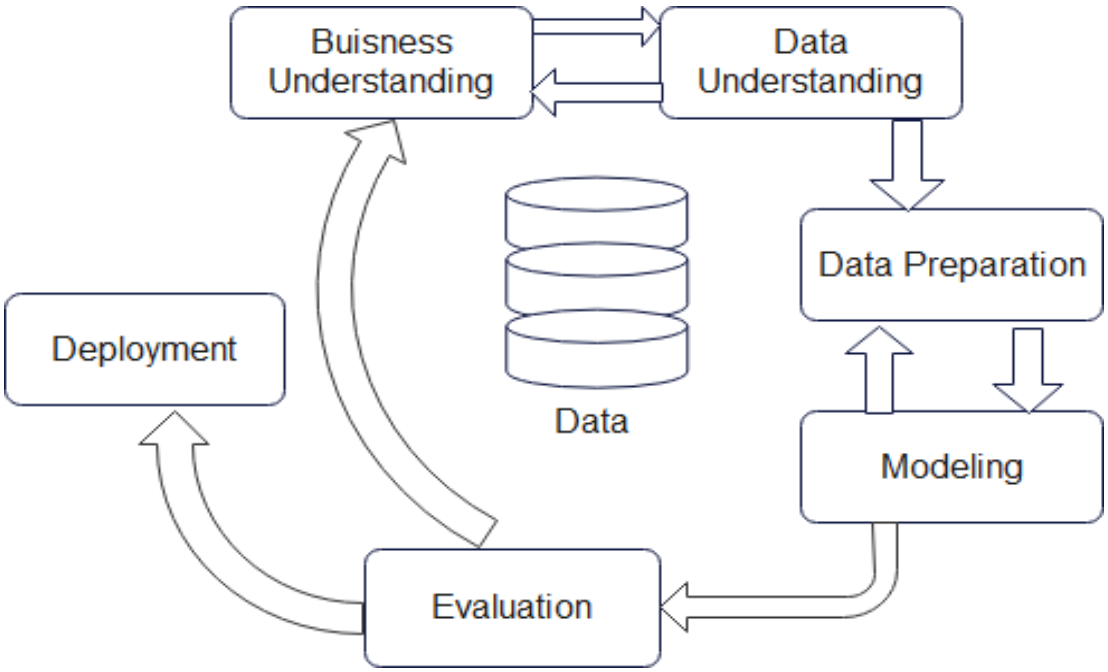



Figure 2.22: CRISP Process Model Adopted from [21]

**Step 1: Business understanding:**

Focuses on comprehending the project's goals and specifications. With the exception of the third task, the remaining three tasks in this phase are fundamental project management procedures that apply to most projects.

**Step 2: Data Understanding:**

In addition to strengthening the basis of business understanding, it concentrates attention on finding, gathering, and analyzing data sets that might assist you in achieving project objectives. Have four tasks:

**Collect initial data**: Acquire the necessary data and (if necessary) load it into your analysis tool.

**Describe data**: Analyze the data and note its surface characteristics, such as data type, record count, or field identifiers.

**Explore data**: Dig deeper into the data. Query it, visualize it, and identify relationships among the data.

**Verify data quality**: How clean/dirty is the data? Document any quality issues.

**Step 3: Data preparation:**

The data is prepared for the subsequent data mining procedure at this step. One of the most crucial and time-consuming parts of data mining is data preparation. Analytics are used to make business choices. However, if the data is erroneous or lacking, your analytics will help you make bad business judgments. Poor analytics result in bad business decisions. Because of this, data from many sources is combined and cleaned up to ensure that there are no duplicate, erroneous, or missing items.

**Step 4: Modeling:**

Various modeling approaches are chosen and used in this phase, and their parameters are calibrated to ideal values. Typically, several approaches exist for different DM issue types. Is the data mining process' analytical hub. Here is where modeling approaches are chosen and applied.in this phase normally dividing the dataset into train, test, and validation sets before actually developing a model is takes place. There are four tasks in this phase:

**Select modeling technique:** Determine which algorithms to try (e.g. regression, neural net).

**Generate test design:** Pending your modeling approach, you might need to split the data into training, test, and validation sets.

**Build model**: As glamorous as this might sound, this might just be executing a few lines of code like “reg = Linear Regression (). fit (X, y)”.

**Assess model**: Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

**Step 5: Evaluation**

The assessment makes sure that the developed data models are accurately compared to the job and chooses the model that is most appropriate. Utilizing the business criteria defined at the project's outset, the outcomes of the preceding processes are assessed. Therefore, the goal of this phase is to examine if the data mining solution solves the business challenge and to see if there is a business reason why a model is flawed. There are three tasks in this phase:

**Evaluate results:** Are the models up to the standards for organizations success? Which one(s) should we approve for the business?

**Review process:** Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.

**Determine next steps:** Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

**Step 6: Deployment**

The chosen model is employed in the deployment step after data preparation, model development, and model verification. The acquired knowledge must be arranged and presented in a way that the end-user can comprehend. Plan deployments, monitoring, and maintenance, presentations, project reviews and documentation make up the bulk of this phase's work.

### 2.2.3 Hybrid process model

A hybrid model was created by combining two or more data mining approaches in order to use the strengths of various classifiers and enhance the classifiers' performance. Hybrid models, or models that incorporate elements of both academic and industrial models, are the result of the evolution of both types of models. A six-step KDP model is one such paradigm. It was created by incorporating the CRISP-DM methodology with scholarly research. This model differs from CRISPDM in that it contains multiple feedback mechanisms, a data mining stage in place of modelling, and generic research description steps [54]. the hybrid process model has several differences with other process model

- Incorporating several more explicit feedback mechanisms (the hybrid model contains seven feedbacks, compared to the CRISP-DM model's three major feedbacks).
- Feedback loops are necessary because decisions and alterations made at one stage may have an impact on changes made in subsequent steps.
- Present each phase in a more generalized, research-focused way.

- Modifications to the final phase since the hybrid approach allows for the application of information from one domain to another. According to [22] the hybrid model has six steps that are explained below:

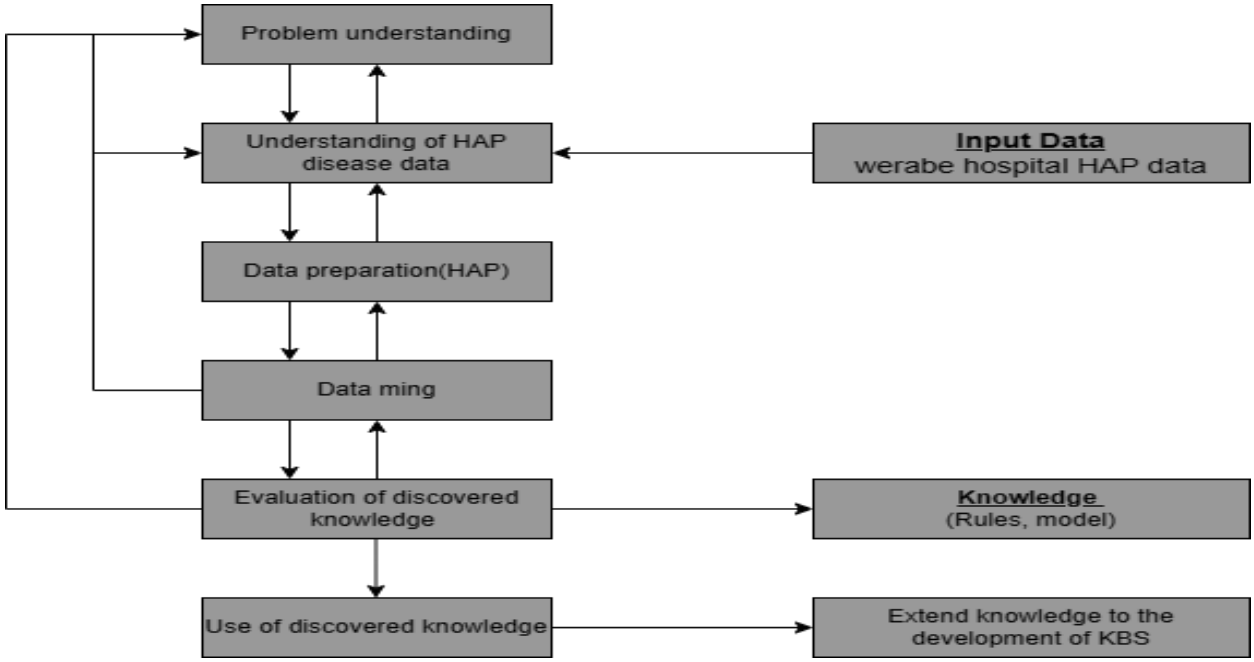



Figure 2.23: Hybrid Process Model adopted from [22]

**Understanding of the problem domain**: During this phase, the highly involved collaborate with the domain experts to specify the project's objectives, identify important players, and discover a suggested solution to the current issue. It also entails picking up vocabulary unique to the domain. A written explanation of the problem is provided, along with its limitations. Project goals are finally converted into DM goals, and the first round of DM tool selection is carried out in preparation for further steps in the process. **Understanding of the data**: Assess which data, including format and amount, will be required based on sample data collected. We'll check for missing values, redundancy, completeness of data, and the plausibility of attribute values. Finally, the usefulness of the data for DM goals will be verified. **Preparation of the data**: Selecting the data that will be the input for the DM methods in the next step is the main goal of this step. In addition to performing correlation and significance tests, sampling and data cleaning are involved.

Data cleaning entails verifying that data records are full and eliminating or adjusting for noise and missing values, among other things. After cleaning, the data can be analysed further using feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes by discretization, and by summarization of data. The results are data that meet the specific

Input requirements for the DM tools selected in the first step. **Data mining** is one of the crucial phases in the process of discovering new information. Finding new information is the primary goal here. Using one of the selected DM tools, the data model was built, and training and testing protocols were created as part of the process of finding new information. After that, testing protocols were used to validate the created data model. **Evaluation of the discovered knowledge**: involves comprehending the findings, determining whether the knowledge gained is new and fascinating, having domain experts interpret the findings, and determining the significance of the knowledge gained. Only models that have been accepted are kept, and the entire procedure is reviewed to determine what more steps might have been done to enhance the outcomes. **Use of the discovered knowledge**: Planning where and how to apply the newly acquired knowledge is the last phase. It is possible to expand the current domain's application area to other domains. A strategy is established to oversee the application of the acquired information, and the project's complete process is recorded. The newly acquired knowledge is finally put to use.

## 2.3 Data Mining Tasks

In general, two types of data mining jobs may be distinguished depending on the objectives of each. These two types of tasks are descriptive and predictive. While predictive data mining jobs use inference on the current data set to forecast how a new data set will behave, descriptive data mining activities define the general qualities of data [17].

Classification, prediction, time-series analysis, association, grouping, and summarizing are a few examples of data mining tasks. All of these activities fall into one of two categories: descriptive or predictive data mining tasks. One or more of the actions listed above can be carried out by a data mining system as part of data mining.

### 2.3.1 Descriptive task

Typically, descriptive data mining activities uncover patterns in the data that describe new, important information from the existing data set. Descriptive data mining jobs describe the fundamental characteristics of data.

#### 2.3.1.1 Clustering

To find data objects that are similar to one another, clustering is utilized. As [24] discussed the act of grouping data into a set of clusters so that each object in a cluster is comparable to another object in the same cluster is known as clustering, and it is a crucial data mining approach.

#### 2.3.1.2 Association rule mining

Association is the process of identifying a connection or relationship between two sets of objects. Association makes links between different parts clear. As the name implies, association rules are straightforward If/Then statements that aid in the discovery of connections between seemingly unrelated relational databases or other data repositories. A process known as "association rule mining" seeks to identify recurring patterns, correlations, or relationships in datasets available in various databases, including relational databases, transactional databases, and other types of repositories [25].

#### 2.3.1.3 Summarization

The generalization of data is summarization. A collection of pertinent data is condensed, producing a smaller set that provides data aggregate information.

### 2.3.2 Predictive task

By drawing conclusions from recent data, predictive mining activities provide forecasts. Prediction is the process of examining the attribute's previous and present states in order to anticipate its future state.

#### 2.3.2.1 Regression

A statistical method for estimating the relationship(s) between a dependent variable and other variables is regression. One or more independent variables (x or predicting variables) and the result variable (y or outcome variable).

#### 2.3.2.2 Classification

Separate data classes or ideas are the process of classification. Finding a model that explains and separates data classes or ideas is the process of classification, with the goal of using the model to predict the class of objects whose class label is unknown [26]. Finding the traits that identify the group to which each instance belongs is the goal of classification issues. Classification may be used to forecast how new instances will behave as well as to comprehend the data that is already available. Decision Trees are a common data mining approach used for categorization purposes**.**

**Decision tree**

Decision tree algorithms are capable of predicting outcomes and finding patterns. The human-generated rules' understandability in large datasets. For instance, understanding knowledge collected in tree form using a decision tree requires less mental effort.one may create rules from the tree to forecast the class for unidentified entries by following the path from the root (the chosen attribute) to leaf (the class label). In addition, decision tree induction's categorization processes are straightforward and quick, and tree creation doesn't require any domain expertise [16].

Decision tree-based classifiers include the methods J48, Random Tree, and REP Tree [26]. The learning and classification processes for decision tree induction are straightforward and rapid. Decision tree classifiers often outperform other learning algorithms in terms of resilience to noise and cheap computing cost for model generation [23].An illustration of a decision tree-based classifier are the J48 algorithm, Random Tree and REP Tree algorithms.

**J48 classification algorithm:** J48 is a WEKA-implemented C4.5 classification method built on decision trees or they are tree-based classifiers in WEKA. A decision tree-based classifier sorts the input instances by moving them down the tree from the root to the leaf node, starting at the top. The anticipated output value for a certain input instance is represented by the value of the leaf node [23].

**Step 1: -** Input the dataset.
**Step 2: -** Check whether all cases belong to the same class.
**Step 3: -** Initialize the tree to form the structure
**Step 4: -** For each attribute z, find the gain ratio value by splitting between z.
**Step 5: -** Supposed that z is an attribute with the highest gain ratio value.
**Stage 6: -** Create the decision node for processing.
**Stage 7: -** Recur on the sub lists obtained by splitting on z best, and add those nodes as children of the node.
**Stage 8: -** Output: The generated decision tree for classification then writes in the form of an if-then rule.

**Random Tree:** The Random Tree Classifier uses bagging techniques to apply the idea of a set. Receives the input characteristic vector, classifies it using each tree in the group of tree predictors, and produces the class label with the highest number of votes using the bagging track to create a sample of random data to create a decision tree classifier model and this classification algorithm is a robust algorithm that can handle noisy data and outliers and it is less likely to over fit the data means it can generalize well to new data and it is stable version of tree based models [30].

**REP Tree:** The rapid decision tree technique known as the tree-based classifier reduces error pruning (REP) was created for regression or decision trees based on the entropy principle and entropy information gain computing, as in algorithm C4.5.
When breaking the matching instances into parts, it is concerned with reducing the mistake brought on by variance and missing data. This REP Tree method uses the regression concept to

build a number of trees through modified iterations before choosing the best tree out of every one created to build a classifier model [24].

**Rule-Based Classification Algorithms: -** The IF-THEN structure, in which the IF component is known as the condition and the THEN part as the action, is used to describe rules. The rule serves as the fundamental knowledge container and unit in rule-based reasoning. Both programmers and subject-matter specialists may quickly understand the IF-THEN rules since they are relatively natural for people. However, it might be challenging to accurately describe a domain expert's understanding of basic rules. By providing a set of rules that may be used to give class labels to new instances based on their features, rule-based classifiers are used for classification. These rules may be developed using domain-specific expertise or they may be automatically learnt using a collection of labeled training data [25].

**Rule :( Condition) =>X**

The foundation of the rule-based reasoning inference engine is the idea that IF the user's data satisfies a rule's requirements, THEN the rule's actions will be carried out. JRip and Partial Decision Tree are two industrial-strength rule induction classification algorithms that are preferred and used in this research because of their simplicity of use and ability to produce the best-performing prediction model, according to [26]. These algorithms are used repeatedly for constricting partial decision trees and generating rules from them.

**JRip:** A propositional rule learner is implemented. JRip presented a Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is an inference and rules-based learner that can be applied to predict elements with propositional rules. The JRip classification algorithm is a straightforward way for extracting rules from data. [24]. JRip (Weka's implementation of the RIPPER rule learner) is a rapid classification algorithm for learning “IF-THEN”, it has the advantage of being a high level and symbolic knowledge representation that contributes to the discoverability of knowledge. Rule learning techniques, like decision trees, are popular because the knowledge representation is simple to understand. The algorithm progresses through four phases [27].

**Phase I: -** In the growth phase, rule is created by greedily adding features to the rule until the rule Meets stopping requirements.

**Phase II: -** In the following prune part; each rule is incrementally pruned, allowing the pruning of any final sequence of the attributes, until a pruning metric is fulfilled.

**Phase III: -** In this stage, each generated rule is further optimized by greedily adding attributes to the original rule and by independently growing a new rule undergoing a growth and pruning Phase, as described above.

**Phase IV: -** Finally, in the selection phase, the best rules are kept and the other rules are deleted from the model.

**PART: -** Partial Decision Tree (PART) is a separate-and-conquer rule learning strategy. The rule induction algorithm producing sets of rules called decision lists which are ordered set of rules. A new instance is compared to each rule in the decision lists, and the item is assigned the group of the first matching rule. PART produces a pruned decision tree using the C4.5 statistical classifiers in each iteration. From the best tree, the leaves are translated into rules. Partial decision tree is a combination of C4.5 and RIPPER [24]. The process of partial decision tree algorithm is presented in detail as follows: -

**Step 1: -** Build a pruned decision tree for the current set of instances.

**Step 2: -** Make a rule from the decision tree to diagnosis Hospital acquired pneumonia diseases.

**Step 3: -** Read off the rule for the largest leaf or the leaf (best) with the largest coverage is converted in to the rule.

**Step 4: -** Afterward, the partial decision tree is discarded which avoids hasty generalization by only generalizing once the implications are known (i.e. all the sub trees have been expanded).

**Step 5: -** Removes the instances from the training collection that are covered by the rule.

**Step 6: -** Continue, creating rules recursively for the remaining instances until none are left or proceeds recursively until no instance remains. This process is called a separate and-conquer strategy. The performance of PART is fast since it does not need any post preprocessing.

## 2.4 Data mining and health care

Finding the best medicine combinations to treat second layer populations that react to particular treatments differently from the afflicted population is one of the current uses and problems of data mining in healthcare. Estimates or views that are obtained as a source of information are known as data. There are several types of data that may be represented in various ways. Research on enormous shared clinical datasets and information-driven analysis are gaining momentum quickly and provide considerable freedoms to enhancing health systems as well as individual care. The healthcare community may benefit from such accessible data by better understanding the underlying causes of disease and, eventually, the outcomes of therapy [28].

IT solution quickens Create a diabetes screening model employing data mining technologies, which reduces the ongoing costs associated with patient management. By facilitating communication, encouraging evidence-based decision-making, utilizing previously stored data, sharing success stories, incorporating e-learning to remote health professionals, using it as a medium to access recent healthcare information, and engaging in data handling and processing activities, different information technologies (Its) can help save lives in developing countries' healthcare systems [29].

Data mining has been increasingly popular in the fields of science, engineering, bioinformatics, genetics, and medicine in recent years. It is a collection of algorithmic techniques for extracting instructive patterns from unstructured data and is used as an illustration in the healthcare industry. It is crucial in combating the data overload in medical informatics and offers a user-oriented approach to novel and hidden patterns in the data. By using extracted patterns, applications can be created to assess the efficacy of medical treatments and identify the behaviors of disease [17].

## 2.5 Attribute selection measure

Because they specify how the tuples at a certain node are to be split, attribute selection measures are also known as splitting rules. The splitting attribute for the provided tuples is determined by

the attribute with the highest score for the measure. Not all characteristics are equally significant when categorizing a given dataset with regard to a certain target class. A good selection will increase the classification accuracy. Attribute selection measure is primarily used to choose the splitting criterion that best divides the provided data division. Because they control how tuples are divided at a certain node, feature selection measures are also acknowledged as splitting rules. In general, this section discusses the Information Gain, Gain Ratio, and Gini Index as three widely used feature selection metrics [19].

### 2.5.1 Information gain

As implied by the name, this measurement is based on the quantity of data required to categorize the data in Partition D at a node N. This signifies that an attribute is a candidate for splitting if and only if it requires less information to categorize the data partition at that node. Additionally, the chosen splitting attribute should lessen the "impurity" in the divisions that it produces [19]. The expected information needed to classify a tuple in D is given by:

$$\text{Info (D)} = -\sum_{i=1}^{m} pilog2(pi) \quad \text{.....2.1}$$

Where pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by |Ci, D|/|D|. Now, suppose we were to partition the tuples in D on some attribute A having v distinct values, {a1, a2… av}, as observed from the training data. If A is discrete-valued, these values correspond directly to the v outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets, {D1, D2…Dv}, where Dj contains those tuples in D that have outcome aj of A.

However, in spite of this much information the partition may be impure so the amount of information that

Would be required to arrive at an exact classification is measured by: -

$$info\ A(D) = \sum_{i=1}^{M} \frac{|Dj|}{|D|} * Info(Dj) \quad \text{.....(2.2)}$$

The term |Dj |/|D| acts as the weight of the jth partition. Info A (D) is the expected information required to classify tuple from D based on the partitioning by A.

The smaller the expected information required, the greater the purity of the partitions.
Information Gain is defined as the difference between original information requirement and the new Requirement. That is,

$$Gain(A) = Info(D) - Info_A\,D \text{ .....(2.3)}$$

**Gain (A**) informs of the gain that would result from branching on A. The anticipated decrease in the amount of information needed as a result of understanding the significance of a. At node N, the splitting attribute is determined to be the attribute A with the largest benefit. This is akin to saying that we want to partition on the attribute that would do the "best classification" such that the quantity of information remaining needed to finish classifying the tuples is as little as possible.

### 2.5.2 Gain ratio

Gain ratio is an improvement on information gain that seeks to address the flaws of information gain. The problem with information gain is that it is more bias towards the selection of the attribute that gives many outcomes.
Gain ratio adds one more value “split information” for overcoming the problems with the information gain measure.

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{Dj}{D} * \log_2 \frac{Dj}{D} \text{ .....(2.4)}$$

$SplitInfo_A(D)$ Represents the prospective data produced by partitioning the training data set, D, into v parts, each of which corresponds to v results of a test on Aspect A. The gain ratio is outlined as follows: -

$$\text{GainRatio(A)} = \frac{\text{Gain(A)}}{\text{SplitInfo(A)}} \text{ .....(2.5)}$$

The attribute with the highest gain ratio is the one that was selected as the splitting attribute.

## 2.6 Performance evaluation

Data mining classification techniques have the capacity to handle a lot of data. Data is categorized based on the training set and class labels and it can predict categorical class labels. Accuracy, True Positive Rate, False Positive Rate, Precision, and F-Measure are often used metrics to gauge how well a predictive model is doing. By displaying how frequently instances of a class, such as class label A, are mistakenly identified as belonging to class A or another class, such as class label B, the confusion matrix aids in understanding how well a classifier performed [30].

**Confusion matrix**

Each instance is classified by the binary decision tree classifier as either True or False, according to one of two predetermined classifications. The right categorization model, which is actually positive and actually negative, may be found in the confusion matrix table along the main diagonal. The second field displays the categorization mistake [31].

Table 2.1: Confusion Matrix of Model Evaluation

| | | Predicted | |
|---|---|---|---|
| | | Negative(N) _ | Positive(P) + |
| Actual | Negative_ | True Negative (TN) | False Positive (FP) |
| | Positive+ | False Negative (FN) | True Positive (TP) |

### 2.6.1 Accuracy True Positive Rate and False Positive rate

Contrary to Predictive Accuracy, the values of the TP rate and FP rate are independent of the proportion of positive and negative classifications [30].

**True Positive rate (TP): -** is the proportion of positive or correctly classified instances as positive or Correct instances.

$$\text{True posetive rate} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \quad (2.7)$$

**The False Positive (FP): -** rate is measures the proportion of negative instances that are erroneously Classified as positive.

$$\text{False posetive rate} = \frac{\mathbf{TN}}{\mathbf{TN+FP}} \quad (2.8)$$

### 2.6.2 Precision, Recall and F-Measure

**Precision: -** It is an indicator of the level of accuracy attained in real prediction. In plain English, it informs us of the proportion of real positive forecasts to all positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{\text{Predictions actually posetive}}{\text{Total predicted posetive}} \quad (2.9)$$

**Recall: -** Recall is the ratio of the total number of correctly categorized positive classes to the total number of positive classes. Or, how much of all the positive classifications we accurately anticipated.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{\text{Predictions actually posetive}}{\text{Total actual posetive}} \quad (2.10)$$

**F-measure / F1-Score: -** is a value between 0 and 1 that represents the harmonic mean of recall and accuracy. Because it is not sensitive to extremely high values, we employ the harmonic mean.

$$\text{F} - \text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \quad (2.11)$$

## 2.7 Knowledge base system

Knowledge-based systems, a kind of artificial intelligence, are computer programs that can execute tasks and make choices based on input from users while attempting to mimic the thought processes of human experts. How does artificial intelligence work? Making the computer system act or think like a human. The expertise of the expert is accessible when a human expert

would not be, enabling information to be made available whenever and wherever it is required. Expert systems get input for decisions from the user interface's prompts or from data files that are kept on the computer [32].

Computer programs called knowledge-based systems use an inference process and a knowledge base to solve issues, provide new information (like a diagnostic), or offer advice. Most systems include a user interface and some level of explanatory functionality. Information-based systems are defined as focused on the acquisition, representation, and application of information that is unique to a given activity, but they also address the enlarged perspectives of such systems made possible by the capacity to apply the same knowledge in many contexts. Knowledge is the most crucial component of any expert system. The precise, superior knowledge that domain expert systems have about certain job domains is what gives them their strength [33].

### 2.7.1 Types of knowledge base system

There are different types of knowledge-based systems. These systems frequently have built-in capabilities for addressing problems, enabling them to comprehend the context of the data they analyze and process and make defensible conclusions based on the information they hold [34].

**Case based system:** These systems develop answers to a problem using case-based reasoning. By looking through historical data from comparable circumstances, this system operates. This involves reviewing past knowledge of similar situations. Based on what it finds, the knowledge-based system provides solutions that were effective in those given situations.

**Classification system:** To determine a data's categorization state, these systems evaluate several types of data.

Expert system: These are a typical kind of KBS that mimic human expert decision-making in a certain sector. Expert systems offer explanations for issues as well as their remedies. They could be applied, for instance, to computations and forecasts.

**Rule based system:** depend on manually created, hard-coded regulations. These guidelines are used to evaluate and alter data in order to get desired results. This can entail applying IF-THEN

rules, which specify that if a user submits a specific request, the system will provide a specific result.

### 2.7.2 Architecture of knowledge base system

A blueprint for an object or system's structure is called architecture. The conceptual model known as the architecture describes the system's structure, behavior, and additional points of view. Additionally, a system's architecture aids in describing the norms, laws, and standards that need to be applied to the related system. Similar to other systems, a knowledge-based system has an architecture that identifies its key elements, essential functions that are carried out by the system, and fundamental tools that support knowledge-based system development [35]. The architecture of knowledge base system is as follow in figure 2.4.

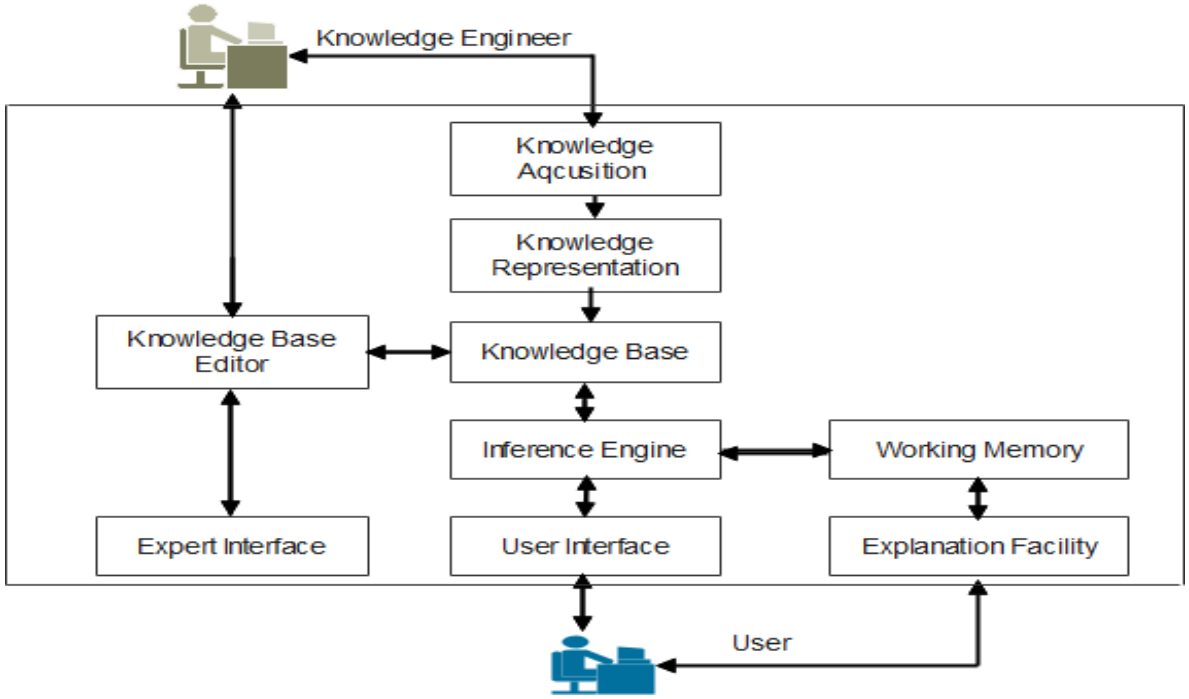



Figure 2.4: Knowledge base system architecture adopted from [35]

**User interface: -** is the point of interaction between a user and a system. The user interface might be either a command line interface (CLI) or a graphical user interface (GUI). Data mining

and knowledge-based systems are being integrated, and the integrator is using a graphical user interface while the knowledge-based system is using a command line interface.

**Inference engine: -** An expert system's brain is the inference engine. It provides a reasoning technique while utilizing the control structure (rule interpreter). The basic objective of the inference engine is to navigate a maze of rules and come to a decision. Backward chaining, one of the most popular inference techniques, is the process of starting with conclusions and moving backward to the supporting evidence. Facts are the starting point of a forward chain, which leads to conclusions.

**Forward chaining: -** working backwards from the conclusions to the facts. In a forward chaining system, the original facts are processed first, and the rules are then applied to fresh data to generate new conclusions.

Backward chaining: is the method of drawing conclusions first, then going back to the data that support them. The hypothesis (or solution/goal) that we are aiming to prove is processed first in a backward chaining system, and we continue looking for rules that would enable concluding that hypothesis.

**Knowledge base: -** The knowledge base consists of the information required to grasp, conceive of, and resolve problems. It is a repository of knowledge in a particular field that the knowledge acquisition component has collected from human experts.

**Explanation facility: -** Giving people explanations so they can comprehend how a system operates and judge whether or not its logic is sound is crucial in a different field.

Knowledge engineer: A knowledge base is built by a knowledge engineer using knowledge from datasets, experts, research papers, theses, or observation. In order to codify and make explicit the rules that a human expert utilizes to solve actual situations, he or she enlists the assistance of human domain experts who collaborate with the knowledge expert. As a result, he or she consults with subject-matter experts while designing, developing, and debugging the knowledge base [36].

**Knowledge acquisition: -** The process of collecting, transferring, and transforming problem-solving expertise from specialists or recorded information sources into a computer program for building or growing the knowledge base is known as knowledge acquisition. In addition, it deals with extracting and displaying human specialists' knowledge. It is the most crucial phase of KBS development. The most challenging part of creating KBS is acquiring knowledge since the expert lacks sufficient understanding of programming and expert system methodologies and finds it challenging to accurately and fully articulate his expertise [37].
The validity, precision, and dependability of the knowledge that is elicited determine how well knowledge-based systems work. The methods of factual and explicit knowledge collection that are frequently employed are document analysis, observation, and interviewing.

**Knowledge representation: -** The process of representing information obtained from the subject matter expert in a usable form, frequently in the form of IF-then-else rules, is known as knowledge representation. The knowledge base is a systematic and structured representation of subject-matter expert knowledge that can be accessed quickly to solve particular problems. The branch of artificial intelligence known as knowledge representation (KR) is concerned with the automated manipulation of knowledge by reasoning programs. While representation is a blend of syntax, semantics, and reasoning, knowledge progresses from data to information to knowledge to wisdom. KR is therefore a field of research in AI that seeks to express information in symbols to make it easier to infer from those knowledge components, hence producing new knowledge elements [32].

According to [37] explanation, knowledge-based systems aim to incorporate the expertise of a human expert (such as a highly qualified doctor or lawyer) into a "computerized consulting service" that does not get tired, bored, or old because such systems preserve and disseminate the knowledge so that it can be used in the future. An expert system uses an embedded reasoning process in its inference engine, or the "thinking" element of the system, to deliver guidance based on its knowledge base. Computer software called an expert or knowledge-based system is created to emulate the decision-making abilities of decision makers or experts in a certain, constrained field of expertise.

A sound knowledge representation method serves as a conduit for achieving human expressiveness, computing efficiency, and a foundation for reasoning. In intelligent systems, there are two common methods for problem-solving: rule-based reasoning and case-based reasoning. Cases reflect specialized information, whereas rules represent the domain's broad knowledge [38]. The following rule-based reasoning methodologies were covered in this study's work.

## 2.8 Knowledge-Based Reasoning Techniques

There are numerous methods of knowledge-based reasoning. Rule-based reasoning, case-based reasoning, ontology-based reasoning, and semantic network reasoning are some of the popular reasoning approaches. The rule-based strategy, which is employed in this study, is the most widely utilized method.

### 2.8.1 Rule Based Reasoning (RBR) Techniques

It is the most important type of legal reasoning. in this type of reasoning a case holding rule is taken and it is applied to a set of fact. One of the most often utilized reasoning paradigms in artificial intelligence (AI) is rule-based reasoning.

The knowledge base, which typically consists of a set of "IF... THEN..." rules representing domain knowledge, and the inference engine, which typically contains some domain-independent inference mechanisms like forward-backward chaining, are the two main parts of the reasoning architecture of rule-based systems. In figure, the overall RBR solution is displayed. An input problem is first matched against the knowledge base's rules to identify any applicable ones. Next, the selected inference mechanism (such as forward or backward chaining) generates intermediate outcomes, and the process is repeated until the desired solution state is attained [39].

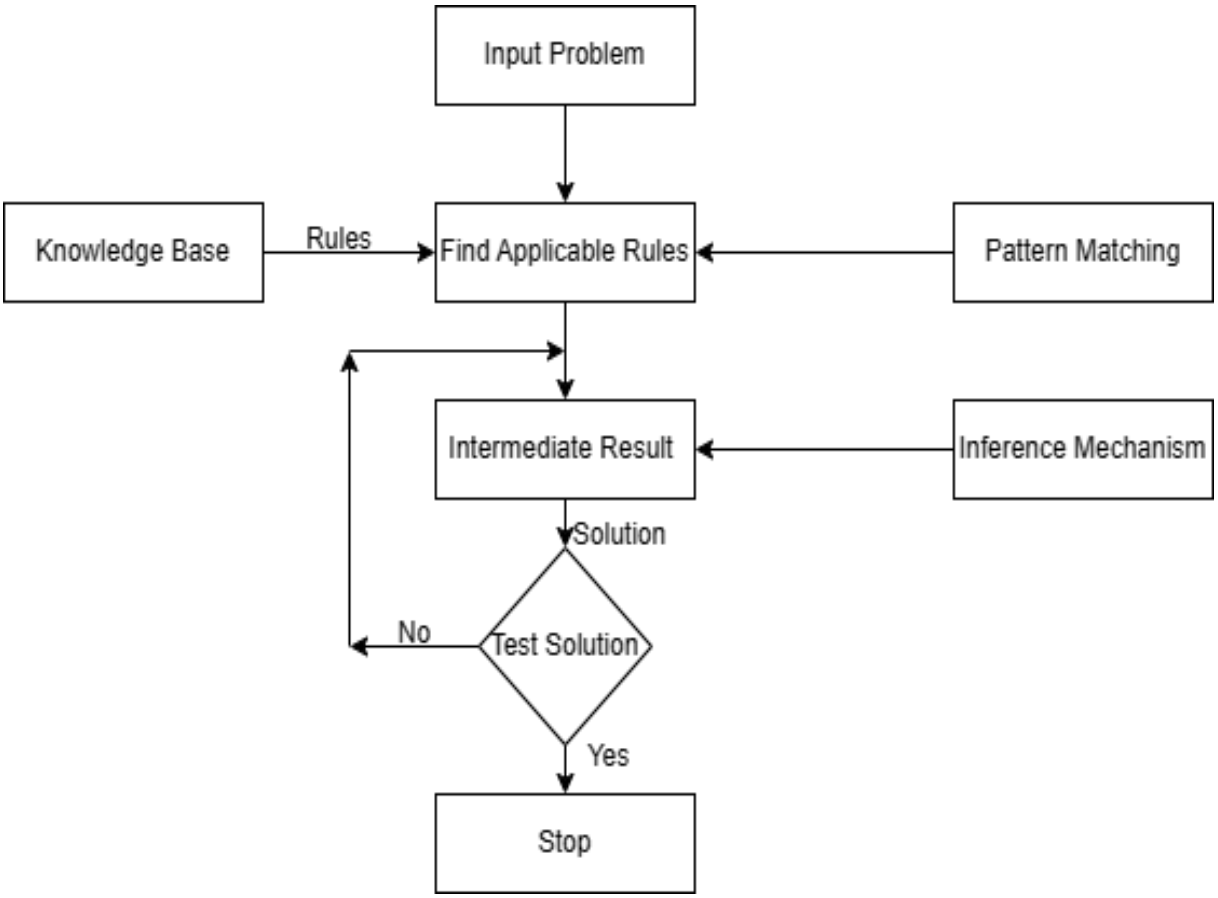



Figure 2.5: Rule Based Reasoning Adopted From [19]

**Rule based reasoning**

In rule-based reasoning, there are two primary inference procedures. The forward chaining and the backward chaining are these. In contrast to the latter, the former is driven by the objectives (conclusions).

**Forward chaining**

Forward chaining is a data-driven method where reasoning starts with previously known information and moves forward using that information until it achieves a conclusion or a goal, or until the full chain has been trained. Working memory for facts is continuously updated in a data-driven control system.

When certain criteria are met, rules in the system reflect potential courses of action that might be taken to hold things in working memory. Actions often include adding or removing things from working memory while forward chaining. Given the working memory, the interpreter of the inference engine manages the application of the rules [40].

**Backward chaining**

Backward chaining, also known as goal-driven reasoning, is the second strategy in rule-based reasoning [40]. Backward chaining rules begin with a hypothesis to be tested and look for information already in existence that meets that hypothesis or a rule that can infer that hypothesis. The condition of a rule becomes a sub-goal that now has to be proven when a suitable rule has been discovered and chosen. The system then searches for rules that can provide these sub-goals. In other words, the backward chain traverses the rule's chain in the opposite way to the forward chaining when conducting a pattern match between the goal and the right side of the rule.

In contrast to backward chaining, which starts with the goal and moves backward by applying inference rules to find the facts that satisfy the goal, forward chaining starts with known facts and moves forward by applying inference rules to extract more data? This process continues until it reaches the goal. The forward and backward chaining approaches are shown in Figure 2.6.

Initial Facts: A, B, C, D, E and Rules: -

1. IF A AND B, THEN F RULES
2. IF C AND D, THEN G RULES
3. IF F AND G, THEN H RULES
4. IF E AND H, THEN I

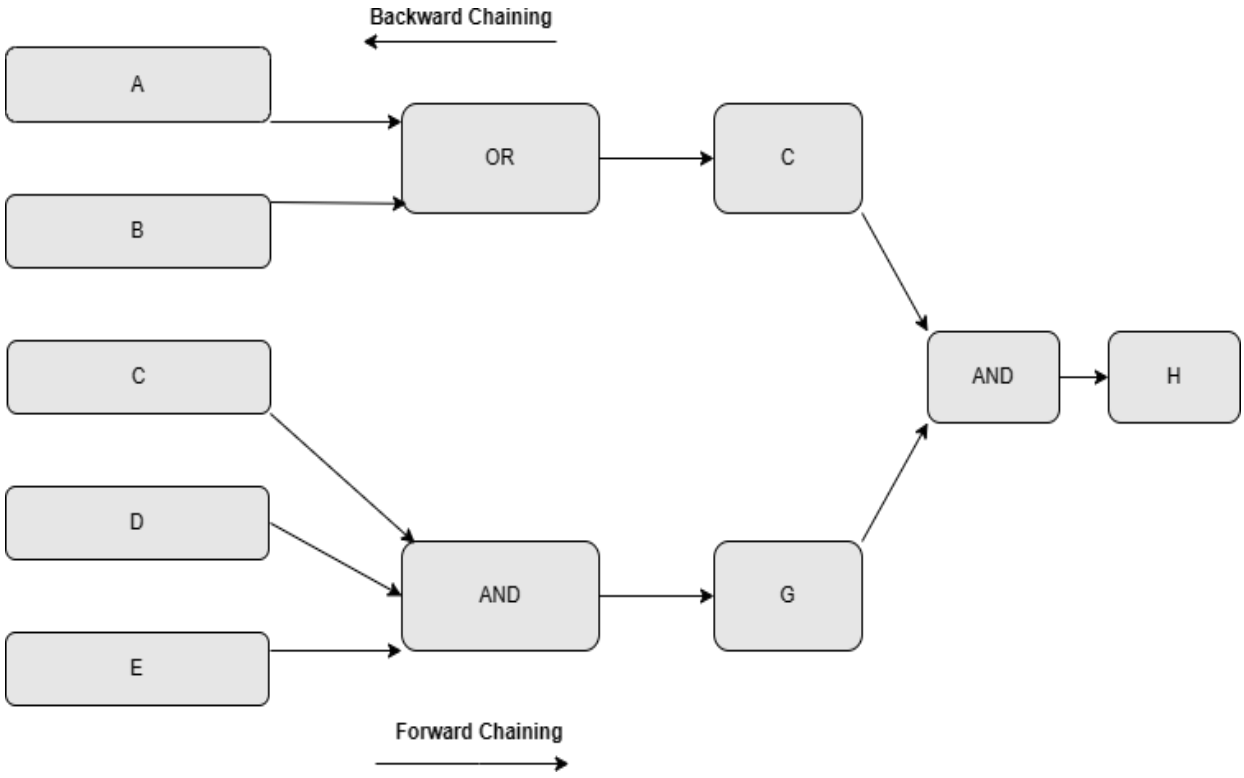



Figure 2.6: Forward and Backward Approaches

### 2.8.2 Case Based Reasoning Technique

Though frequent and crucial to human cognition, case-based reasoning (also known as analogical reasoning) has only lately become a significant way of thinking. By locating and modifying like difficulties kept in a library of prior experiences and problems, this style of thinking includes solving new problems. The case library (stored representations of prior experiences or issues) is the foundation of the CBR's reasoning architecture and a cycle of inference. Finding and retrieving examples from the case library that are most pertinent to the situation at hand (input) and adapting the obtained cases to the current input are crucial processes in the Case Based Reasoning (CBR) inference cycle.

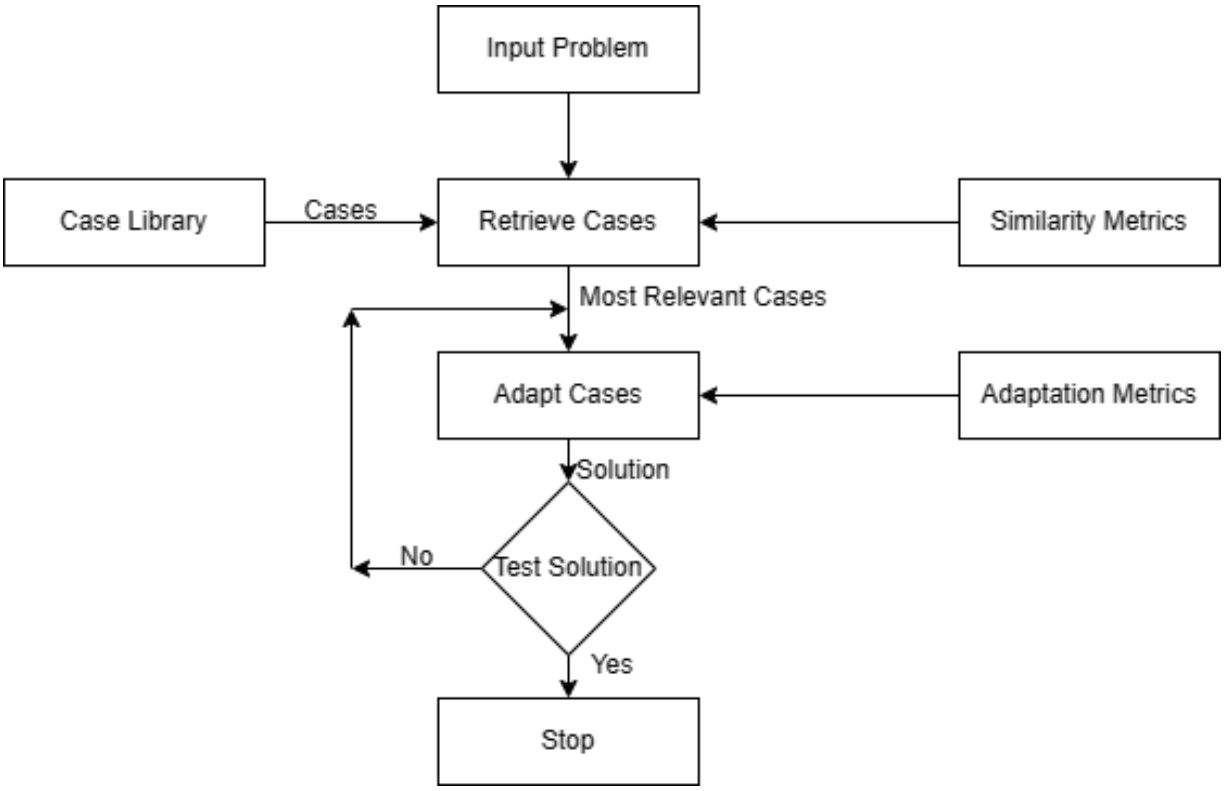



Figure 2.7: Case Based Reasoning

## 2.9 Knowledge Base System Implementation Tool

Knowledge representation (KR) is the area of artificial intelligence that deals with reasoning systems manipulating knowledge automatically. While knowledge develops from data to information to knowledge to wisdom, representation is a combination of syntax, semantics, and reasoning. In order to make it easier to deduce new knowledge elements from existing knowledge components, researchers in the AI subject of knowledge representation want to express information in symbols.

A set of programs comprising several software tools and instructions that aid in the creation of a knowledge-based system is called a KBS tool. A range of alternatives are available for defining information and applying knowledge to decision-making in knowledge-based system programming tools. The goal of these technologies is to quickly construct a knowledge-based system [15].

These multifunctional programming languages can be used to develop non-AI applications. To create KBS, programming languages such as Prolog is utilized [59]. One programming language that is ranked higher than LISP is Prolog, which is regarded as high-level since it offers the programmer better functionality.

A popular programming language in artificial intelligence research is Prolog. It has multiple more logical characteristics in addition to a purely logical subset known as "pure Prolog". Programming in declarative logic is done with Prolog. Program execution begins with a query against the relations that define the program's logic. The prolog interpreter generates the answer based on the stated conditions after we specify the rules, facts, and conclusion while implementing a problem's solution.

The computer language SWI-Prolog is employed for this investigation. A contemporary portable Prolog compiler that complies with ISO includes the following features: availability, flexibility, portability, and the ability to integrate with other programs [41].

## 2.10 Related Works

Around the past few decades various investigations have been carried out in various institutions and research facilities all around the world. This section investigates related works done on diagnosis and treatment of Hospital acquired pneumonia (HAP) disease.

**Vikas Chn** [2]**,** developed a novel transfer learning based approach for pneumonia detection in chest X-ray images.The objective of this study is to make pneumonia detection easier for both beginners and specialists. This study used the pretrained architectures, AlexNet, DenseNet121, Inception V3, Google Net, and ResNet18 trained on the Image Net dataset, to extract features in this framework using the transfer learning approach. These Characteristics were transferred to the classifiers of the corresponding models, and each architecture's output was gathered. Ultimately, the researcher utilised an ensemble model that surpassed all other models by utilising all five pertained models. After that, the investigator suggested an ensemble model that integrates outputs from every pretrained model, surpassing the performance of individual

models and attaining the highest level of performance in pneumonia identification. Using unseen data from the Guangzhou Women and Children's Medical Centre dataset, the ensemble model achieved an accuracy of 96.4%.

**Zafaret al** [3]**,** the main objectives of this study was to set out to determine and quantify the risk factors that distinguish pneumonia from upper respiratory infection. The main source of data was collected from World Health Organization. The study used 259 cases of pneumonia with 187 cases of cough and cold those attributes are taken as a controlled variable among children 5 years of age at a large tertiary care hospital in gilgit Pakistan. By analysing clinical parameters and determining determinant variables by logistic regression, the researchers were able to create a clinical score that could be used to predict the diseases. In the multivariate logistic regression analysis, lack of immunization adjusted odds ratio (AOR) = 1.54, 95% CI 1.0, 2.3, previous history of pneumonia (AOR=1.77, 95% CI 1.16, 2.7), younger age (AOR) for each preceding month in children aged up to 59 months=l. Ol, 95% CI 0.99, 1.03) and malnutrition (wasting) (AOR=2.2, 95% CI 1.0, 5.23) were revealed as important risk factors for pneumonia.

**Elina Naydenova** [42], has analyzed power of data mining in the diagnosis of childhood pneumonia. The source of data analyzed here was originally collected as clinical study. 1581 participants were Gambian children aged 2–59 months. Various features were collected for each case. The full dataset consisted of 57 features (clinical characteristics), including measurable clinical variables (e.g., white blood cell count, neutrophils, haemoglobin, etc.), observational clinical characteristics (e.g. sleepiness, sternal in drawing, cough heard, etc.) and conventional vital signs (e.g., respiratory rate (RR), heart rate (HR), oxygen saturation (Osat), etc.).
Pre-processing was done on the initial data. Imputation was used to handle the significant number of missing values (up to 42% for some features) by removing features and cases that made up less than 85% of all the entries. The remaining missing values were imputed using feature median value.in this study three classification techniques namely Random Forest (RF), Logistic Regression (LR) and Support Vector were applied (SVM). Some differences were

observed between the classifiers SVM and RF performed comparably whereas LR achieved better with five features LR delivered 84.6% RF 71.8%.

**Enshishu Tse** [5]**,** developed Data mining result-based hybrid knowledge-based system for pediatric community acquired pneumonia diagnosis and treatment system the case of bale robe general hospital. The source of data was pediatric ward of bale robe general hospital. And total records of 2990 dataset with 22 attributes were collected and the data was preprocessed using data mining preprocessing tools followed Cross Industry Standard Process. The researcher developed predictive model using J48, REPT tree classifier and random tree while the descriptive model was built using Make Density Based Clustering Algorithm, Expectation Maximization and K-means clustering algorithms. The proposed system achieved an accuracy of 96.33%. With domain expert evaluation the result of the user acceptance test verified 91.5%.and consequently, an overall performance of prototype 99% precision and 94.1% f-measure achieved respectively.

**Yuanyuan Li** [9], the objectives of this study is evaluating the diagnostic performance of Deep Learning models in detection and classification of pneumonia using chest X-ray (CXR) image. The main source of data for the study were public media, web of science and Google scholar in order for retrieving all studies for implementing a deep learning algorithms for distinguishing pneumonia patients from healthy using chest x-ray images (CXR).and bivariate linear mixed model to pool diagnostic estimate is used such as sensitivity (SE), specificity (SP), positive likelihood ratio (PLR), and diagnostic odds ratio (DOR). In test set about 11292 CXR images used reported from fourteen studies and 3357 images from patients with pneumonia and 7447 image datasets from healthy control were included. Regarding the pathogenic type of pneumonia about 781 images from patients with bacterial pneumonia and 597 images were from patients with viral pneumonia.

The main preprocessing task performed in this study are extracting regions of interest (ROI) and their features, preprocessing of images and then finally feature based classification of the diseases. And finally, the researcher achieved 95% confidence interval in discriminating pneumonia Chest X-ray (CXRs) from controls this shows that DL indicated high accuracy

performance in classifying pneumonia from normal Chest X-ray (CXRs) radiographs and in distinguishing bacterial from viral pneumonia.

Table 2.2: Summary of Related Works

| **Author** | **Title** | **Methods and approaches** | **Result and Evaluation** | **Finding** |
|---|---|---|---|---|
| Vikas Chn (2020) | A novel transfer learning-based approach for pneumonia detection in chest X-ray images. | Decision tree, neural networks and ensemble approaches such as random forest. | Accuracy of 96.4% achieved with an ensemble approach (random forest) | Complexity of diagnostic process<br>Expert knowledge is not collaborated.<br>Uses limited data set lead to over- fitting<br>Prevents for achieving higher accuracy. |
| Zafaret al (2017) | Comparison of 'cough and cold' and pneumonia: risk factors for pneumonia in children under 5 years. | Logistic regression, manual knowledge acquisition. | Achieved accuracy of **95%** using logistic regression | Uses domain expert knowledge only.<br>Doesn't consider important attributes<br>Expert knowledge is not collaborated. |
| Elina Naydenova (2016) | The power of data mining in the diagnosis of childhood pneumonia | Logistic regression (LR), support vector machine (SVM)& random forest (RF) | The accuracy of the model is 86.4% using Random Forest | Expert Knowledge is not included<br>Used limited data set excluding important patient data this leads to prevent for achieving a higher accuracy. |

| | | | | |
|---|---|---|---|---|
| Enshishu Tse (2021) | Data mining result-based hybrid knowledge-based system for pediatric community acquired pneumonia diagnosis and treatment system the case of bale robe general hospital. | Decision tree using different Classification algorithm Technique. | REP Tree classifier and K-means clustering algorithm has highest accuracy of 96.33%. | Development of KBS is not considered with other classification algorithm Such as (ANN, SVM, Rf, and LR).<br>His work is focused on community acquired pneumonia which is less sevior and has less mortality rate than HAP.<br>He doesn't make severity assessment of the diseases that helps expert to diagnose the diseases early. |
| Yuanyuan Li (2020) | Accuracy of deep learning for automated detection of pneumonia using chest X-Ray images. | Introduce Milord II Modular language for KBS as an appropriate tool and to develop KBS. | Deep learning algorithm | The accuracy of the result is not specifically reported as numerically, rather it is concluded as being a promising result than a real human expert. |

## Summary of related works

As sown in the table 2.2 above many studies globally and locally attempt to diagnose pneumonia using different data mining and machine learning techniques. And some of the researcher used knowledge-based system so as to diagnose and treat pneumonia. Studies conducted by [2] try to develop novel transfer learning approach for detecting pneumonia diseases using chest X-ray images as source of data but the researcher used limited number of data sets this hinders from obtaining good accuracy and leads to an over fitting in the general result. Studies carried out by [3] make analysis and experiment by using only knowledge obtained from experts and the researcher doesn't consider relevant attributes for achieving the objectives. Studies carried out by [42] In order to obtain some of the critical parameters, the study should be expanded in both the analysis of larger and higher datasets for increasing and obtaining a good accuracy and the creation of suitable point-of-care instruments e.g. detection algorithms for lung sounds via stethoscope. The study conducted by [5] even though he is developing a KBS the development of knowledge base system doesn't considered with other important classification algorithms. He doesn't evaluate the diseases severity, which would enable a professional to make an early diagnosis. He focused on community acquired type of pneumonia which is less sevior than other type of pneumonia. And he used limited amount of data set since data mining requires huge amount of data for obtaining best result and doesn't specify important parameters for diagnosing and making treatment recommendations. Studies conducted by [9] the researcher tries to show that Deep learning (DL) had high accuracy performance in the classifying of pneumonia from normal Chest X-Ray (CxR) radiographs and also in distinguishing bacterial from viral pneumonia however the assessed Deep learning (DL) diagnostic accuracy does not reflect clinical practice, because the researcher do not compared with the performance of health-care professionals.

Therefore, by considering the limitation and recommendation of the above related works this study aims to design a knowledge-based system that might assist domain specialists in the diagnosis and treatment suggestion of hospital acquired pneumonia diseases Additionally, using data mining classification algorithms like J48, JRip, PART and Random Forest because

It is the most powerful classifier for removing an over fitting. On the dataset that was acquired from werabe hospital, rule-based reasoning will be built. Furthermore, the specialist domains will provide the tacit knowledge. And finally, the data mining classification result collaborate with knowledge of expert for making diagnosis and treatment of the disease.

# CHAPTER THREE

# RESEARCH METHODOLOGY

The goal of research methodology is to understand which research methodologies and procedures will be used for the goal of data collection, preparation, organization, analysis and visualization, and interpretation for the study. Research methodology is a road map that demonstrates how the researcher will study from beginning to end. In general, the researcher summarizes the overall research methodology, research process models, and algorithms in this chapter.

## 3.1 General research approach

There are several data mining techniques available today, and they are suitable for use in a variety of healthcare settings, including clinical decision support systems, patient care quality improvement, and health care research and development. Data mining techniques are employed in both developed and developing nations to forecast the occurrence of various health care issues, such as epidemics, community and facility-based screening programs.

The overall research technique in this study has been conducted using the design science research methodology. A technique known as "design science research" defines and operationalizes research when a recommendation or artifact serves as the intended objective. Theories may be improved by the creation or assessment of artifacts through design science research, which are categorized as constructs, models, methodologies, and instantiations.

The first justification for the use of DSRM in this study is that it usually entails the development of an artifact known as KBS as a way to enhance the state of practice for the diagnosis and appropriate treatment of the chosen Hospital acquired pneumonia diseases (HAP). Second, as the established KBS not only solves the current issue but also serves as a guide for the advancement of theories, it may be able to close the gap that now exists between theory and practice [41].

In line with earlier research, Peffers et al [41] propose a DSRM that offers a process model for design science research technique in order to convey the DSR. According to the authors, this process model supports researchers and is an effective method for carrying out research within the framework of design science. We defined the DSRM process model in this section. The research method can be employed in a variety of ways depending on the kind of topic being studied, as well as the study purpose and beginning point. The problem-centred initiation serves as this research's entry point in light of these beginnings. As depicted in Figure 3.2, the components of the DSRM process model are

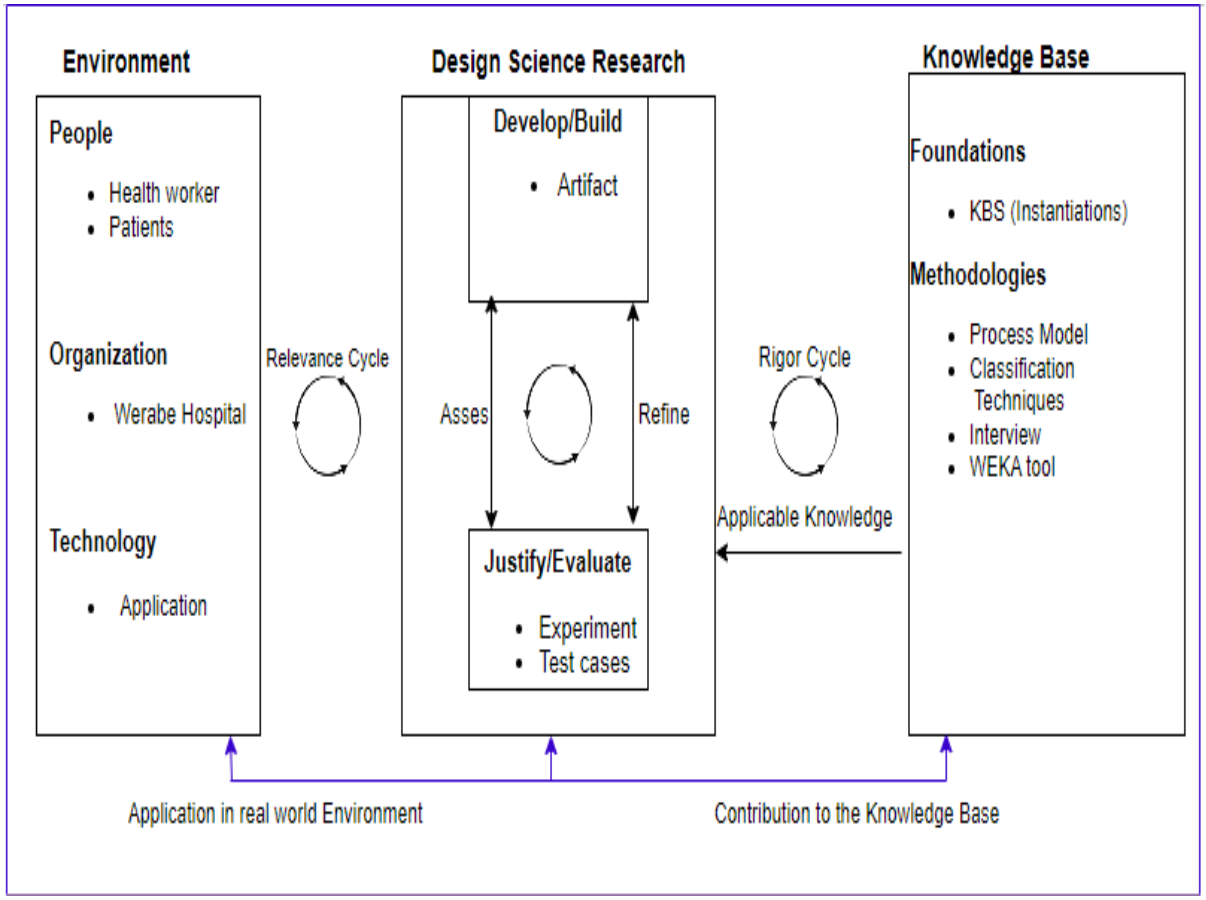



Figure 3.1: Design Science Research Framework Adopted From [14]

The environment in which the problem is being observed that is, the location of the phenomena of interest to the researcher is referred to as the environment in the above image. The people, the organization, and its technology make up this environment.

Considering the noted demands of the organization and when it comes to issues that the researcher is interested in, design science research can help build and develop artifacts and reinforce the body of current knowledge [41]. The environment where a researcher may identify which hypotheses or artifacts have been used or generated by other scholars is known as the knowledge base. Additionally, it designates the site where the materials needed to create new research and artifacts are gathered. It is made up of approaches and foundations.

## 3.2 Design Science Research Methodology

To achieve the specified goals, the design science research process model was employed. As a problem-solving strategy, it creates a new, intentional artifact to address a generalized problem type and assesses its usefulness for resolving that kind of problem. Design Science Research Methodology, consists of six actions, the first four of which offer potential starting or ending points for DSR studies. The input from the DSRM's prior work, which is available, for example, in the form of an article, can serve as the starting point for a DSR investigation. Problem centred initiation, or defining and identifying the challenges to be solved through an artifact, is the starting point for this study [41] the eps of design science research are described below

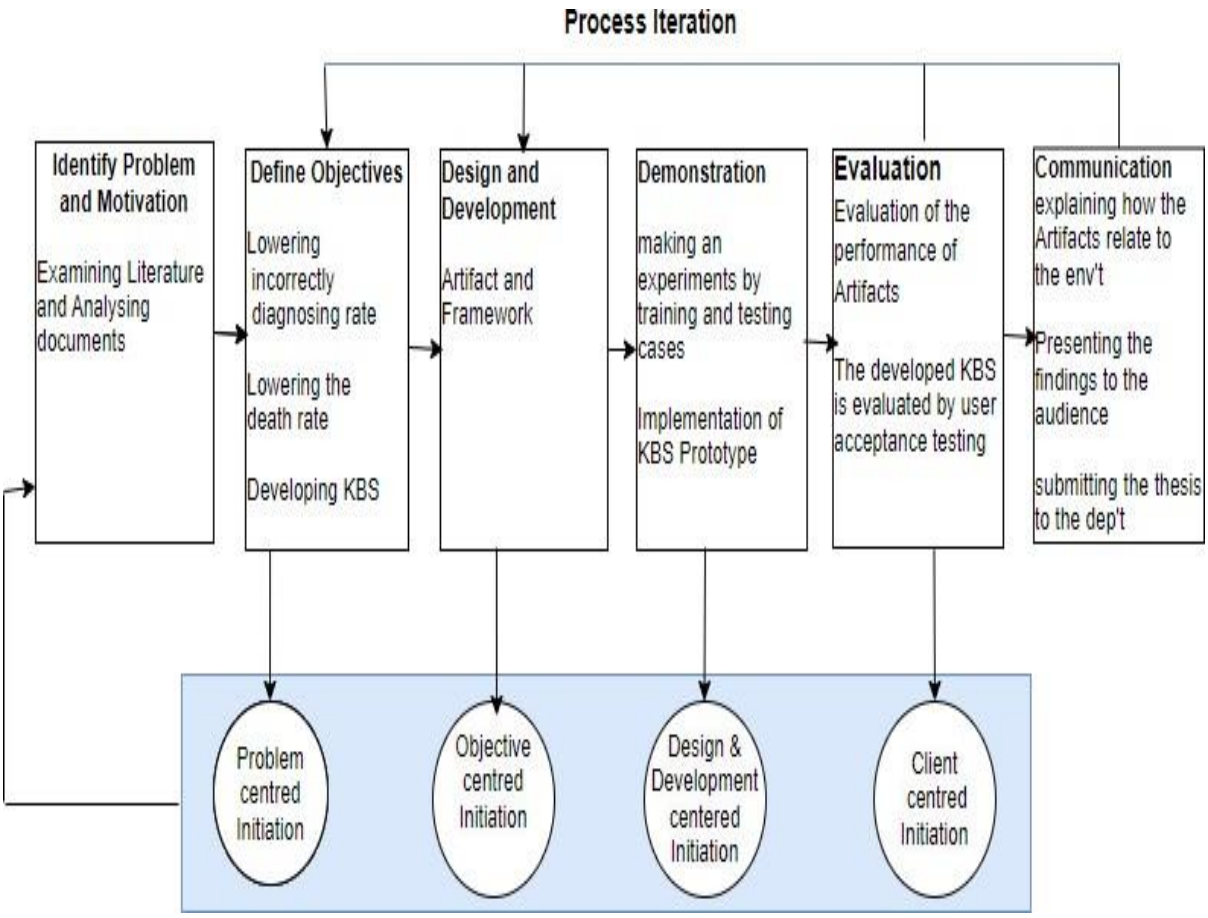



Figure 3.2: Design Science Research Process Model Adopted from [14]

### 3.2.1 Problem identification and Motivation

The process model of the Design Science Research Methodology (DSRM) was selected for this investigation. DSR is a systematic approach to problem solving that yields novel, creative and pertinent information systems solutions within a certain domain [22]. The problems in this study identified through literature review, various relevant documents analysis and interviewing domain experts. The problem identified here is, the absence of an intensive care unit for severe ill patients in the medical facility, a shortage of trained healthcare personnel, too weak quality of services, and a lack of medical services and awareness are some of the issues linked to the diseases. To fix such problems the researcher initiated to develop classification model with

knowledge base system for diagnosis and treatment recommendation of Hospital Acquired Pneumonia (HAP) that assist experts.

### 3.2.2 Define the objective

Determining the objectives is the next stage in the design science research technique, following problem identification. The primary goal of the study is to develop classification model with knowledge-based system that will assist professionals in promptly diagnosing hospital acquired pneumonia (HAP) and recommending the best course of action that is right treatment. The created knowledge-based system decreases patient waiting times for diagnoses, lowers the rate of incorrect diagnoses, and ultimately lowers the death rate.

### 3.2.3 Design and development

This phase involves developing or creating the artifact (KBS) that will aid in problem solving. This step defines the intended functionality, suggested architecture, and development of the item. Furthermore, this step includes the following descriptions of the data sources, knowledge modeling, knowledge representation, and data mining techniques applied to the prediction model. These artifacts could be models, instantiations, structures, or procedures [22].

#### 3.2.3.1 Knowledge discovery process model

The knowledge discovery process model is made up of a number of techniques that are used to give good coverage, starting with the comprehension of the problem domain, followed by the understanding, preparation, analysis, and mining of the data, evaluation of the knowledge found, and application of the information found. Because it combines the best aspects of the CRISP and KDD methodologies to identify and explain several explicit feedback loops that aid in achieving the study objectives, the hybrid data mining process model is used. Additionally, explicit information is obtained from Werabe Hospital. Expert knowledge and data mining results are utilized in the development of KBS. As a result, to create the final framework for the suggested system, the researcher combines an expert knowledge acquisition process with the data mining process model[43].

#### 3.2.3.2 Proposed framework

The suggested framework for this study is a model created with the stated goals that is developing classification model with knowledge base system for diagnosis and treatment of HAP in mind. Three stages are involved in the building of a framework: **knowledge-based systems**, which are created by combining data mining results with expert knowledge, **manual knowledge acquisition** from domain experts, and **automatic knowledge acquisition** through data mining from the dataset. The proposed system's framework is shown in Figure 3.3.

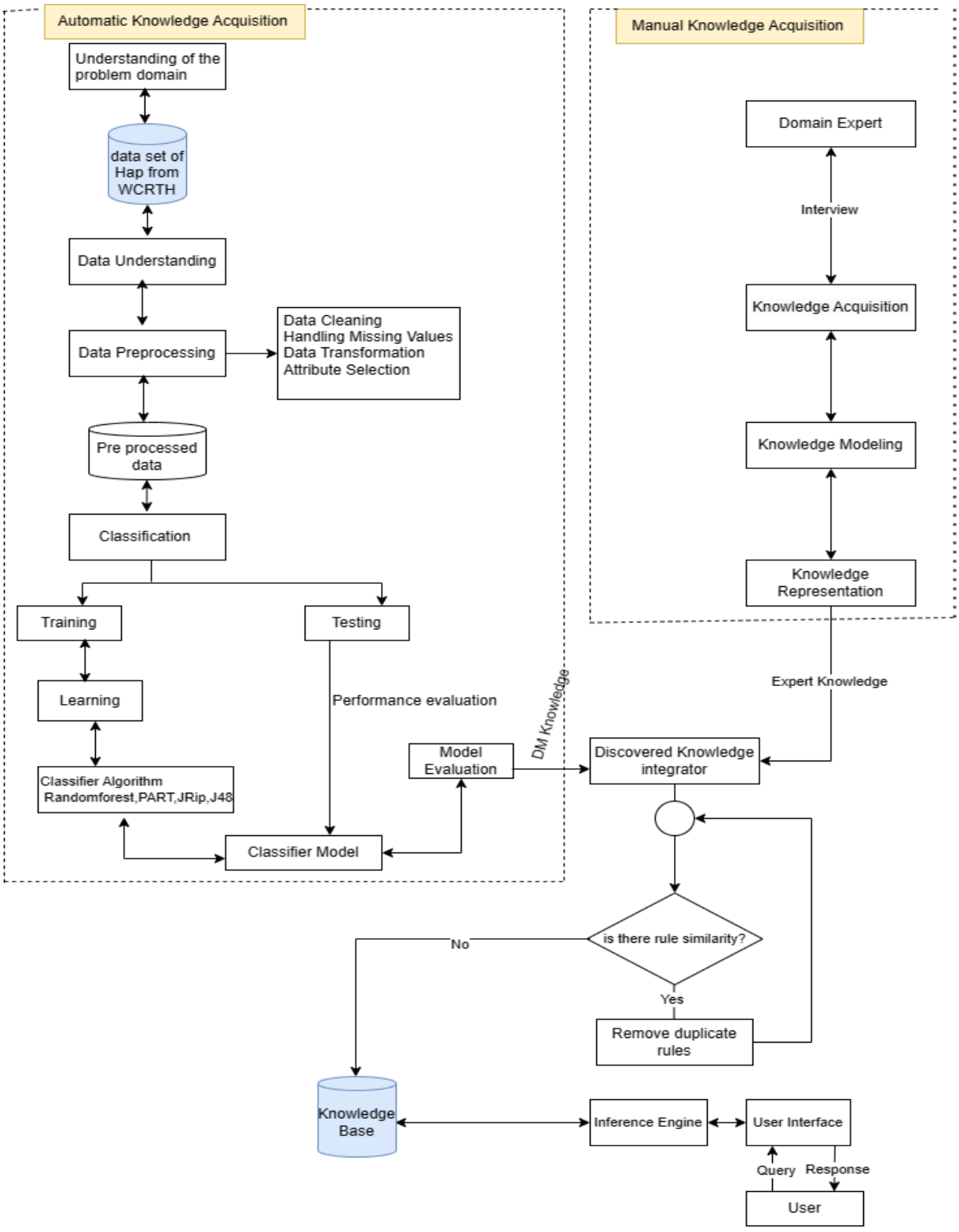



Figure 3.3: Framework of Proposed System

**Understanding of the problem domain: -** The researcher understands the problem domain by communication through interviews with relevant domain experts. This first phase entails remaining with them to comprehend the issue, establish the project's objective, and find out about the problem's existing solution. Based on his observations, the researcher looks into a variety of issues, including patient care methods, disease diagnosis, and therapy recommendations. To learn from the stored data, a more effective classification method from the start is crucial. Reviewing of related works and discussions with the above-identified stakeholders help to understand an overall problem area.

**Knowledge acquisition: -** knowledge acquisition focuses on relevant activities such as gathering important knowledge analyzing the document, identifying important outlooks focusing on the disease that is needed for the study. To acquire this knowledge the researcher used both primary and secondary source of data. The primary data were collected from medical experts who work in the health center by using semi structured interview. The knowledge which is obtained from the experts is used to understand the area and to develop a KBS. The secondary sources of datasets were acquired from pneumonia records admitted to the hospital. And an automatic knowledge is acquired by using data mining classification algorithm.

**Automatic Knowledge Acquisition: -** One of the key challenges in creating a knowledge-based system is gathering information from subject-matter experts. The main cause is that a human expert will not have enough understanding about intelligent systems, which will prevent him or her from effectively and accurately communicating what they know. Automatic knowledge acquisition is required to address the challenges associated with traditional information acquisition and to enhance knowledge-based systems. For this study an automatic knowledge is acquired by using classification techniques such as J48, JRip, and Random Forest and PART tree algorithms.

**Dataset: -** An important dataset for this study is dataset of Hospital Acquired Pneumonia (HAP) collected from Werabe referral hospital. And the researcher has collected seven years of dataset and records it into Microsoft Excel for further preparation and getting ready for input to WEKA

3.8.2 data mining tool. Following the necessary data's collection, it is examined for accuracy, consistency, missing values, and usefulness with respect to the data mining objectives.

**Data understanding: -** After understanding the problem to be addressed the next step is analysing and understanding available data. The primary source of the dataset used in this study was actual patient data obtained from werabe Referral Hospital pneumonia follow-up units. Data understanding includes understanding the nature of the collected data that is important for next phase of the study. Total number of records is 3244 with 21 attributes

**Data preprocessing: -** After the data is collected the next step is preprocessing of the data that includes correction of values that is unavailable and pre-processing of data to remove undesired attributes or minimize noise for mining results. It involves making sure that all data values for the associated attributes are accurate and consistent. The data were produced for knowledge discovery as a result. The researcher converts all nominal data using the Microsoft Excel 2016 tool. And in order to increase the performance of the model different preprocessing technique were applied such as data cleaning; missing value handling, data transformation and attribute selection techniques are performed.

**Data cleaning: -**To obtain a consistent dataset and improve the chances of effective data mining. Making sure that data records are complete, removing or compensating for noise, filling in missing information, and other similar activities are examples of this type of operation.

**Missing value handling: -** Before developing data mining algorithm fixing the values that are missing in the dataset is required. In this study, missing values for numerical characteristics are handled by the mean, whereas missing values for nominal parameters are handled by the parameter's mode. However, data duplication reduces data quality, producing a poor result. In this study, duplicate datasets are removed in order to reduce dataset size. Some of the potential problems that might arise from missing values include bias arising from differences between complete and incomplete data, loss of efficiency, and difficulties in processing and interpreting the data [44].

To prevent this problem, it is crucial to handle missing value and there are different techniques are applied ignore the instances containing missing values, filled the missing values manually or with a constant, and imputation methods. For this studythe missing value is treated as follows after the mean and mode imputation techniques were applied.

The mean of all known values for each column or feature is used to fill in the missing values.

Let $X_i^j$ be the $J^{th}$the missing attribute of the be the $I^{ih}$instance, which is imputed by:

$$X_i^j = \sum_{K \in I(\text{complete})} \frac{X_k^j}{n|I(\text{complete})|}$$

Where $I$ (complete) is a set of indices that are not missing in $X_i^j$ and $n|$ (complete)| is the total number of instances where the $j^{th}$ attributes are not missing. Phases to impute the missing values with the most frequent corresponding feature values.

**Step 1:** Count the number of missing values for each value.
**Step 2:** Find the highest occurrence of each attribute
**Step 3:** Impute the missing values with the highest corresponding feature values.

**Data Transformation: -** The process of data transformation involves transforming the data into formats suitable for data mining. Among the processes used in data transformation are aggregation, normalization, generalization, and discretization. The researchers in this study, discretization strategy and expert advice were used to transform the distinct value of attributes. Discretization is capable of not only improving the readability and understand ability of the dataset for the users but also improve predictive accuracy for the induced model. Discretization is one of the most influential data pre-processing tasks to handle numerical attributes in data mining, and it divides continuous or numerical data into categorical or nominal data

**Attribute selection: -** one of the most important tasks in data mining is selecting essential attributes that have an impact in the model to be built. Which involves selecting a subset of the

original feature spaces based on their ability to discriminate in order to enhance the dataset's quality? The selection techniques for the Gain Ratio attributes were used for the feature ranking in this study. Since Gain ratio is a normalization of information gain that considers the post-partition entropy of the probability distribution subset the researcher has used the gain ratio attribute selection measure as discussed in section 2.5.2 [19].

**Classification: -** Several classification methods are employed to create the predictive model for the chosen diseases. These algorithms include decision tree (PART, J48, and JRip), logistic regression, support vector and neural network and the one with good accuracy among them has been employed to work with the KBS's expert knowledge.

**J48 classification algorithm: -** J48 is a WEKA-implemented C4.5 classification method built on decision trees or they are tree-based classifiers in WEKA. A decision tree-based classifier sorts the input instances by moving them down the tree from the root to the leaf node, starting at the top. The anticipated output value for a certain input instance is represented by the value of the leaf node [23].

**Step 1:** Input the dataset.

**Step 2:** Check whether all cases belong to the same class.

**Step 3:** Initialize the tree to form the structure

**Step 4:** For each attribute z, find the gain ratio value by splitting between z.

**Stage 5**: Create the decision node for processing based on their gain ratio value through decreasing order.

**Stage 6:** Output: The generated decision tree for classification then writes in the form of an if-then rule.

**PART classification algorithm:** The separate and conquer method is used by this rule induction classifier, which creates a rule, eliminates the instances it covers, and keeps producing rules recursively for the examples that remain until none are left. The general procedures of PART concerning this research are as follow.

**Step1:** decision tree is constructed based on the current set of instances.

**Step 2:** a rule is built from the decision tree to identify the class type (severe, mild or moderate).

**Step 3:** To create a classification, rule the leaf with the most coverage is chosen.

**Step 4:** Once the rule is chosen, remove the decision tree.

**Step 5:** If the specified instances do not delete the subsequent rule that is the exact same as the preceding rule, then eliminate the instances covered by the rule.

**Step 6:** Continue doing the tasks until the rules are satisfied.

**JRip classification algorithm**: is a straightforward way for extracting rules from data [45] JRip (Weka's implementation of the RIPPER rule learner) is a rapid classification algorithm for learning "IF-THEN", it has the advantage of being a high level and symbolic knowledge representation that contributes to the discoverability of knowledge. Rule learning techniques, like decision trees, are popular because the knowledge representation is simple to understand. The algorithm progresses through four phases [45].

**Step 1:** In the growth phase, rule is created by greedily adding features to the rule until the rule meets stopping requirements.

**Step 2:** In the following prune part, each rule is incrementally pruned, allowing the pruning of any final sequence of the attributes, until a pruning metric is fulfilled.

**Step 3:** In this stage, each generated rule is further optimized by greedily adding attributes to the original rule and by independently growing a new rule undergoing a growth and pruning Phase, as described above.

**Step 4:** Finally, in the selection phase, the best rules are kept and the other rules are deleted from the model.

**Random Tree: -** The Random Tree Classifier uses bagging techniques to apply the idea of a set. Receives the input characteristic vector, classifies it using each tree in the group of tree predictors, and produces the class label with the highest number of votes using the bagging track to create a sample of random data to create a decision tree classifier model

**REP Tree: -** The rapid decision tree technique known as the tree-based classifier reduces error pruning (REP) was created for regression or decision trees based on the entropy principle and entropy information gain computing, as in algorithm C4.5. When breaking the matching instances into parts, it is concerned with reducing the mistake brought on by variance and

missing data. This REP Tree method uses the regression concept to build a number of trees through modified iterations before choosing the best tree out of every one created to build a classifier model [45].

**Model Evaluation:-**after the model is built the next step is evaluating the performance of the model using metrics such as Accuracy, True Positive Rate, False Positive Rate, Precision, and F-Measure to gauge how well a predictive model is doing [30].

**Manual knowledge acquisition: -** In order to get implicit information regarding the diagnostic and treatment recommendations for the diseases from specialists in the field, both organized and informal discussions were used. Since learning from an expert that is deeply rooted and unique in the domain expert's mind is one of the investigation's specific goals. The investigator conducts interviews with specialists to acquire insights and suggestions for possible solutions to enhance current procedures and obstacles.

Many experts took part in the interview process for this study. The experts were questioned about their prior knowledge in order to fully comprehend the domain area and to determine the key input elements that affect the diagnosis and treatment of Hospital Acquired Pneumonia (HAP). The selection criteria were determined by the educational degree level and the professions/specializations.

Table 3.1: Domain Expert Biography

| **Specialization** | **Work experience** | **Work area** |
|---|---|---|
| Medical doctor | Six years | Werabe Hospital |
| Nurses | Three years | Atat Hospital |
| Internist | Intern | Wolkite university referral Hospital |
| Health officer | Eight | Werabe Hospital |
| Pediatrician | Three | Werabe Hospital |

**Knowledge modeling: -** Following the information gathered from the experts, the researcher proceeded on to model knowledge using the decision tree knowledge modelling technique. The decision tree is an effective tool for capturing rules or more explicitly logical phrases, which is the reason why this modelling technique was chosen. Additionally, a decision tree can be simply transformed into if-then rules that computer systems can use and comprehend.

**Knowledge representation: -**this is the process of converting domain knowledge in to computer understandable form using knowledge representation techniques. Since the knowledge that the researcher acquire from data mining classification technique are in the form of rules and the knowledge that are acquire from document analysis and domain experts' interview about diagnosis and treatment of Hospital Acquired Pneumonia (HAP) are decision trees and procedures which are easy to convert to rules, the researcher used production rule (rule-based knowledge representation method) that is IF-THEN is used. Knowledge which was elicited has been transformed into rules in the form of IF-THEN statements, where conclusions are drawn from the THEN clause only if the IF clause's requirements are met.

**Combining knowledge: -** after converting the domain knowledge in to computer understandable form or rules the next step is combining the knowledge the knowledge originated from rules extracted from classification algorithm and from domain experts in order to have a knowledge base.

The main importance of combining the two-source knowledge is for increasing the system diagnostic accuracy, to acquire a new pattern from the data mining, and to acquire additional patterns from the expert, that are not generated from the data mining for instance the treatment recommendation.

The extracted rules are either kept in two different knowledge bases or integrated into a single knowledge base. When a user inputs a query and it is stored in a different knowledge base, the inference engine examines each knowledge base independently and returns the answer. When a user inserts a question and it is stored in a single knowledge base, the inference engine consults the combined knowledge base to find the answer [46]. Because the knowledge base is simple to use, the extracted rules from this study have been consolidated and represented as a single

knowledge base. The next stage after incorporating the knowledge is to verify rule duplication and eliminate the redundant rules. The duplication of the rules is checked using the pseudo code that follows.

The pseudo code used to check the rule redundancy

1. Knowledge from expert and data mining result
2. Check for rule redundancy
3. If all rules are different store the rules into the knowledge base
4. If a similar rule has appeared remove redundant rules and store them in the knowledge base

**Knowledge base: -**includes the information required for understanding, formulate, and resolve issues. It is a warehouse for domain-specific knowledge that the knowledge acquisition module has obtained from human experts. The knowledge generation rules are represented using semantic nets, frames, logic, and other tools. All pertinent data, rules, cases, and relationships that the expert system uses are kept in one place the knowledge base.

**Inference engine:** is an expert system's brain. It employs the rule interpreter control framework and offers a reasoning process. It interprets the rules by processing and analysing them. It is employed in the process of matching antecedents between firing rules and user replies. An inference engine's primary job is to navigate a maze of rules in order to reach a conclusion the inference engine's purpose is to search the knowledge base for relationships and facts and then use that information to generate suggestions, answers, and predictions just like a human expert would. The right facts, interpretations, and rules must be located by the inference engine and correctly assembled [39]. There are two types of inference procedures that are frequently used: forward chaining and backward chaining. The technique of working backward from conclusions is known as "backward chaining."

**User interface: -** serves as a centre for communication between users and the system. Command line interface (CLI) or graphical user interface (GUI) are two possible forms of the user interface. During the process of integrating data mining with knowledge-based systems, the integrator's graphical user interface and the knowledge-based system's command line interface

**Explanation facility: -**It gives the user information in response to queries posed by the system. This feature helps provide clarity when responding to the system's inquiries. The method asks questions in order to identify diseases such as internal parasite, mastitis, enteritis, pneumonia, foot mouse disease, blackleg, and pneumonia. If the user is unclear about the question, they can obtain an explanation.

#### 3.2.3.3 Implementation tool

Version 3.8.2 of the Waikato Environment for Knowledge Analysis (WEKA) is used for pre-processing, uncovering hidden knowledge, and comparing various classifiers on pre-processed datasets. SWI-Prolog 8.2.4 is used to represent rules in the knowledge base and to develop the Knowledge-Based System. Waikato Environment for Knowledge Analysis (WEKA) version 3.8.2 is used for pre-processing, to extract hidden knowledge, and to compare different classifiers from a pre-processed dataset obtained from Werabe comprehensive teaching referral hospital. The researcher created a user interface using Java Net Beans as well.

### 3.2.4 Demonstration

A demonstration is the design science research model (DSRM's) fourth stage. The created artifacts are applied in this phase to address the issue. Both simulation and experimentation can be used to carry it out. The created artifact, or KBS, for the diagnosis and recommended course of treatment of hospital acquired pneumonia illnesses illustrates the diagnosis outcome in this investigation. Clear understanding of how to apply the artifact to the problem's solution is necessary [47]. The suggested structure was then put into practice utilizing tried-and-true development techniques and resources. During this stage, expert knowledge and the outcomes of data mining were integrated to create KBS.

### 3.2.5 Evaluation

This step involves measuring the behaviour of the generated artifact for problem-solving and assessing how effectively it fulfils its anticipated environmental value [30]. By creating test cases, the researcher has compared the artifact performance results with the specifications needed to solve the problem or achieve the desired performance. User acceptability testing

techniques and system performance metrics are used to assess the created KBS. Feed cases are used in system performance testing to see how well the proposed KBS prototype diagnoses problems and repeatedly suggests treatments. In contrast, the process of performing user acceptance testing involves crafting questions specifically for the users.

### 3.2.6 Communication

The design science research model (DSRM) process ends with this phase of the process. The researcher can explain how the artifact relates to the environment in this step. The study findings are presented to the audience and submitted as a thesis document to the department. Depending on the organization's needs, the created artifact can be implemented.

# CHAPTER FOUR
# DATA UNDERSTANDING AND PREPARATION

In this chapter, data understanding and preprocessing phase are briefly described. To comprehend the health domain, the researcher has been collaborated closely with the pediatrician in Werabe Comprehensive Teaching Hospital. The main goal of this study is to fill the gaps in the current system by developing classification model with KBS that should be used in diagnosing and treatment of Hospital Acquired Pneumonia (HAP) disease by assisting domain experts. This KBS was developed by collaborating the data mining results with knowledge acquired from domain experts. Data mining is proven to extract hidden knowledge from a large collection of the dataset.

## 4.1 Understanding of the Problem Domain

In addition to performing professionally, this phase requires having a complete understanding of the issue and the field being studied. The goal of data mining is determined by the kind of problem that needs to be solved, so before starting the actual activity, we should be able to clearly define the problem and have a firm grasp of the dataset that will be used in the data mining.

In this study, domain experts were interviewed, and supporting sources included a survey of various literatures on data mining applications in healthcare, particularly pulmonary disorders. In order to learn more about hospital acquired pneumonia, pertinent materials were thoroughly examined and a discussion with domain experts was held. The researcher was inspired to create a knowledge-based system by combining data mining findings (automated knowledge acquisitions) with expert knowledge (manual knowledge acquisitions)[48].

## 4.2 Understanding of the Data

The next stage is data understanding to identify the accessible dataset after the problem has been defined. The results of data mining and knowledge discovery are significantly influenced by the number and quality of available datasets.

The primary goals of this phase are to gather the target dataset and determine the kind, quantity, and format of the datasets. The completeness, redundancy, missing values, plausibility of feature values, and other issues are examined in datasets. The last phase involves validating the dataset's efficacy in relation to data mining technology goals. Even though the hospital is starting encoding of patient data lately recorded data is encoded so it needs recording of the patient data that is important for further analysis for inputting to WEKA3.8.4data mining tool. The total size of the initial datasets in excel format is 230KB.

The datasets used for diagnosing the diseases have been collected from Werabe comprehensive teaching hospital found in central Ethiopia region, Ethiopia, from 2007 up to 2015 e.c with a total number of 3244.From this data twenty were independent attributes and one dependent attributes (including three class labels). The exam results of one patient are represented by a piece instance in the dataset which are collected from the discharge summary and physical examination. The registered datasets include admission information, delivery information, symptoms, and laboratory results of the patients as listed below.

**Age: -** This is age of patients who admitted werabe referral Hospital. From those admitted patients the dataset indicates that persons from age 2 years to 10 and comparatively elder ages 50 and above are highly affected groups.

**Sex: -** This is the sex of patients who admitted werabe Hospital and the collected data set indicate that males are more affected than females that is about 67.9% of males and 30.24% females.

**Temperature: -** The precise degree of heat or cold that a body experience is temperature. A clinical thermometer is used to measure body temperature. and represents a balance between the heat produced by the body and the heat it loses. In the collected dataset, the value of body temperature was scored in degrees Celsius (°C).

**Chills: -**Feeling of coldness accompanied by shivering which may arise with or without fever. Felt extremely cold. They typically accompany fever and mostly appear at the onset of an

infection Caused when the outside temperature is felt to be low; the body produces heat through involuntary muscle contractions.

**Fever: -** is a high body temperature brought on by an elevation in the hypothalamic temperature set point. The highest limit of a normal temperature in humans is not universally accepted; instead, many sources give numbers that fall between 37.2 and 38.3 °C (99.0 and 100.9 °F) [13].

**Malaise: -** is a term for a general feeling of discomfort, illness, or fatigue that has no clearly identifiable cause.

**Loss of Appetite: -** One of the main signs of serious illnesses is appetite loss, which leads to weight loss. Protein-energy malnutrition (PEM) is a major factor in the unfavorable consequences of these circumstances. Although they have significant adverse effects, pharmaceutical therapies aimed at stimulating hunger are not very effective. Thus, nutritional therapy seems to be the most sensible course of action in order to address deficient nutrition. Nevertheless, there is currently no defined standard methodology for the administration of nutritional supplementation in malnourished, critically sick medical inpatients, and clinical trial data indicating benefits are scarce [41].

**Nausea: -** is a general feeling of uneasiness and disquiet that can occasionally be mistaken for the want to throw up. It is not painful, but if it persists, it can be a crippling symptom that causes discomfort in the back of the throat, chest, or belly.

**Vomiting: -**Is forceful expulsion of the content of the stomach when irritate it via mouth or sometimes the noise, also known as emesis.

**Dyspnea: -** Dyspnea, also called shortness of breath, is the subjective experience of breathing that is made up of several feelings with different levels of intensity.

**Chest pain: -**Chest pain can manifest in various forms, varying from a subtle discomfort to a sharp stab. It can feel searing or crushing at times. Sometimes the pain starts in the jaw, goes up the neck, and then radiates down one or both arms or to the back. You can have achiness or tightness or your chest could feel as though something is pressing down on it [49].

**Shortness of Breath (SoB): -** is a feeling that the lungs aren't getting enough air. You may have tightness in your chest, a sensation of gasping for oxygen, or an increased effort to breathe.

**Decreased Blood Pressure: -** is the decreased pressure in the circulatory system, which is frequently evaluated to aid in diagnosis because it is directly correlated with the heart's force and rhythm as well as the artery walls' diameter and elasticity.

**Rigors: -**A severe cold feeling and shivering that is followed by a fever, usually with profuse perspiration, especially at the beginning or peak of the illness occur during high fevers and are associated with conditions.

**Dullness to percussion: -**is the quiet, muted tone that is produced when percussion is applied to a solid body part, such as the lung. It denotes thicker tissue, like zones of consolidation or effusion2. Chest percussion that is dull indicates that there is less air in the chest because of soft tissue or fluid.

**Increase vocal fremitus (VF): -**Auscultation or palpation can be used to identify vocal fremitus (VF), which is the voice transmitted to the chest wall. Pleural effusion, pneumothorax, and airway obstruction reduce VF and enhance it.

**Abnormal white blood cell: -** A blood test to determine the quantity of white blood cells in the blood is called a WBC count. Leukocytes are another name for WBCs. They aid in the defense against infections.

Typical ranges for total WBC are 8,000/µL to 20,000/µL or microliter. It is linked to a higher likelihood ratio when the white blood cell count is low (less than 8,000/µL) as opposed to elevated (more than 20,000/µL) [14].

**Temperature: -**temperature is the precise degree of body heat or cold. A clinical thermometer is used to measure body temperature, which is equilibrium between the heat the body produces and the heat it loses.

**Respiration: -**It can be expressed more formally as the number of movements that indicate inspiration and expiration in a given amount of time, or as the number of breaths a newborn

takes in a minute. The respiratory rate is often calculated by counting the number of times the chest rises in a minute while taking breaths. The purpose of respiratory rate measurement is to establish if a person's breathing is normal, abnormally high, or abnormally low. A newborn at rest should breathe between 30 and 60 times per minute (bpm) [13].

Table 4.1: Attribute Description

| No | Attribute Name | Data-Type | Domain Values |
|---|---|---|---|
| 1 | Age | Numeric | [1-150] |
| 2 | Sex | Nominal | {male, female} |
| 3 | Chills | Nominal | {yes, no} |
| 4 | Fever | numeric | {yes, no} |
| 5 | Malaise | Nominal | {yes, no} |
| 6 | Loss of appetite | Nominal | {yes, no} |
| 7 | nausea | Nominal | {yes, no} |
| 8 | Vomiting | Nominal | {yes, no} |
| 9 | Chest pain | Nominal | {sharp, stabbing} |
| 10 | Shortness of breath (SoB) | Nominal | {yes, no} |
| 11 | Decreased Blood Pressure | nominal | {systolic, diastolic} |
| 12 | Rigor | Nominal | {yes, no} |
| 13 | Dyspnea | Nominal | {yes, no} |
| 14 | Tachypnea | Nominal | {yes, no} |
| 15 | Dullness to percussion | Nominal | {normal, flat or dull, hyper resonant} |
| 16 | Increased vocal fremitus | Nominal | {yes, no} |
| 17 | White blood cell (WBC) | Numerical | [2560/µL to 32, 000/µL] |
| 18 | Temperature | Numerical | [34°C to 38.9°C] |
| 19 | Respiration | Nominal | <30bpm,30bpm to 60 bpm >60 bpm |
| 20 | Target class | Nominal | {severe, mild, moderate} |

## 4.3 Data Preprocessing

A trained machine learning algorithm's ability to generalize is always significantly impacted by data preprocessing. Unfortunately, because of their enormous size, numerous resources, and gathering process, raw data are very susceptible to missing, noise, and inconsistency techniques. The most important part of the data analysis process is preprocessing, which entails creating the final dataset and contains the data that will be entered into data mining software in the phase that comes after. Preprocessing data is therefore a crucial step in improving data efficiency, getting high-quality outcomes, and learning valuable new information from it. Data transformation, attribute selection, and missing value imputation are some of the techniques used in data pretreatment for this study [19].

### 4.3.1 Missing Value Handling

One typical issue that might impede the data analysis process is missing data. Values that were overlooked in the dataset will reduce a classification's accuracy and have an impact on the models the classification algorithm generates. There are 3244 records in total in the dataset gathered from Werabe hospital, and 20 characteristics. Six characteristics in this dataset are missing: Tachypnea 16(1%), Fever 63(2%) Nausea 37(1%), LoAppetite 2(0%), Dyspnea 14(0%) and malaise 18 (1%). Table 4.2 shows, name of the attribute, number of missing values and imputation techniques. Based on this, the mean values of numeric the missing values were filled in using an attribute that includes missing data. In dealing with a categorical attribute, the most frequent value, or mode, was utilized to fill in any missing values.

Table 4.2: Missing Value Handling

| Attribute Name | Data Type | Number of Missing Values | Imputation Method |
|---|---|---|---|
| Tachypnea | Nominal | 16(1%) | Mode |
| Fever | Numerical | 63(2%) | Mean |
| Nausea | Nominal | 37(1%) | Mode |
| LoAppetite | Nominal | 2(0%) | Mode |
| Dyspnea | Nominal | 14(0%) | Mode |
| Malaise | Nominal | 18(1%) | Mode |

It entails substituting the mean of all known values for a feature (numeric attribute) in the class to which the instance with the missing attribute belongs for any missing data for that feature and modes of categorical attributes or features.

Table 4.3: Sample Data After Handling Missing Value

| 1: age Numeric | 2: sex Nomina | 3: Chest Pain | 4: Chills | 5: nausea | 6: fever | 7: Vomitting | 8: LoAppetite | 9: Dyspnea | 10: SofBreath | 11: DtoP Nominal | 12: InVoFr | 13: AbWBC Nominal | 14: Tachypnea | 15: Rigor | 16: Malaise | 17: DecBP | 18: Resp Nominal | 19: Temp Nominal | 20: Class Nominal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22.0 | M | stabbing | yes | Yes | No | Yes | No | No | Yes | Hyper resonant | No | Leukopenia | Yes | Yes | Yes | Systolic | Normal | Low | Severe |
| 18.0 | F | stabbing | yes | Yes | yes | No | Yes | Yes | No | Dull | Yes | Leukopenia | Yes | Yes | Yes | Systolic | Normal | Normal | Moderate |
| 35.0 | M | Sharp | yes | Yes | No | Yes | Yes | Yes | Yes | Dull | Yes | Leukopenia | Yes | Yes | No | Systolic | Low | Low | Severe |
| 38.0 | M | stabbing | No | Yes | yes | Yes | Yes | No | Yes | Hyper resonant | Yes | Leukopenia | No | Yes | Yes | Systolic | Normal | High | Mild |
| 44.0 | F | stabbing | No | Yes | yes | Yes | No | No | Yes | Normal | No | Leukocytosis | Yes | Yes | No | Diastolic | High | Normal | Severe |
| 51.0 | F | stabbing | yes | Yes | No | No | Yes | No | No | Normal | Yes | Leukocytosis | No | No | Yes | Systolic | Low | Normal | Mild |
| 57.0 | M | stabbing | yes | Yes | No | Yes | Yes | No | Yes | Dull | No | Leukocytosis | No | No | No | Diastolic | Normal | Low | Severe |
| 64.0 | M | Sharp | No | No | yes | No | Yes | Yes | Yes | Normal | Yes | Leukocytosis | Yes | Yes | No | Diastolic | Low | Normal | Moderate |
| 30.0 | M | stabbing | yes | No | No | No | No | No | No | Hyper resonant | Yes | Leukocytosis | No | No | Yes | Systolic | Normal | Normal | Mild |
| 23.0 | M | Sharp | No | No | No | No | Yes | Yes | No | Normal | No | Leukopenia | Yes | No | No | Systolic | Normal | Low | Moderate |
| 32.0 | F | stabbing | yes | Yes | No | Yes | Yes | No | Yes | Normal | Yes | Leukocytosis | Yes | No | Yes | Systolic | Low | High | Severe |
| 19.0 | M | stabbing | No | No | yes | No | No | Yes | No | Normal | Yes | Leukopenia | Yes | Yes | Yes | Systolic | Low | High | Moderate |
| 27.0 | F | Sharp | yes | Yes | No | Yes | Yes | No | Yes | Hyper resonant | Yes | Leukocytosis | No | Yes | No | Systolic | Low | Low | Severe |
| 55.0 | M | stabbing | No | No | No | Yes | No | Yes | No | Normal | No | Leukopenia | Yes | No | No | Systolic | Low | Normal | Severe |
| 42.0 | M | stabbing | yes | No | yes | Yes | No | Yes | No | Normal | Yes | Leukocytosis | No | Yes | Yes | Systolic | High | Normal | Mild |
| 40.0 | F | Sharp | No | Yes | No | Yes | Yes | No | No | Hyper resonant | Yes | Leukopenia | No | No | Yes | Diastolic | High | Normal | Severe |
| 33.0 | F | Sharp | yes | No | yes | No | Yes | No | No | Hyper resonant | Yes | Leukopenia | Yes | No | Yes | Systolic | Low | Low | Mild |
| 55.0 | M | stabbing | No | Yes | No | No | Yes | Yes | No | Normal | Yes | Leukopenia | Yes | Yes | Yes | Systolic | Normal | Normal | Severe |
| 46.0 | F | Sharp | yes | No | No | Yes | No | No | No | Hyper resonant | Yes | Leukocytosis | No | Yes | No | Diastolic | Low | Low | Mild |
| 67.0 | M | Sharp | yes | Yes | yes | No | Yes | No | No | Hyper resonant | Yes | Leukopenia | Yes | No | Yes | Systolic | High | Normal | Moderate |
| 56.0 | M | Sharp | No | Yes | yes | No | Yes | No | No | Normal | Yes | Leukocytosis | No | Yes | No | Systolic | Low | Low | Severe |
| 54.0 | F | stabbing | No | No | No | Yes | No | No | Yes | Hyper resonant | No | Leukopenia | Yes | Yes | No | Systolic | Normal | Normal | Severe |
| 23.0 | M | stabbing | yes | No | yes | No | Yes | Yes | No | Normal | Yes | Leukocytosis | No | Yes | Yes | Diastolic | High | High | Severe |
| 22.0 | F | Sharp | yes | Yes | No | No | Yes | No | Yes | Hyper resonant | Yes | Leukopenia | No | Yes | Yes | Systolic | Low | Low | Severe |
| 35.0 | F | Sharp | No | Yes | No | Yes | Yes | No | No | Normal | Yes | Leukopenia | Yes | No | Yes | Diastolic | Normal | Low | Moderate |
| 57.0 | F | stabbing | yes | Yes | yes | No | No | Yes | Yes | Normal | Yes | Leukopenia | Yes | No | Yes | Systolic | Low | Low | Moderate |
| 60.0 | M | Sharp | No | Yes | No | Yes | Yes | Yes | No | Hyper resonant | Yes | Leukopenia | Yes | Yes | Yes | Systolic | High | Low | Mild |
| 64.0 | F | Sharp | yes | No | yes | No | Yes | No | No | Hyper resonant | No | Leukopenia | Yes | No | Yes | Systolic | Low | Normal | Moderate |

### 4.3.2 Data transformation

As part of the preprocessing stage, data transformation is essential to guaranteeing the quality of the data before analysis. The process of altering the format or organization of data is known as data transformation. Data discretization techniques were used to reduce the number of possible values for a given continuous attribute value by partitioning the range of attributes into intervals [50]The original data is consequently reduced and made simpler by substituting a small number of interval labels for an attribute's continuous value. This results in a clear and simple-to-use understanding of mining outcomes. Additionally, a domain expert's proposal was added to change the unique value of attributes.

The following attribute values are transformed such as Respiration Rate, decreased blood pressure, decreased white blood cell and temperature.

Table 4.4: Data Transformation [51]

| Attribute Name | Range | Transformed Feature |
|---|---|---|
| Temperature | < 36.5 ℃ | Low |
| | 36.5 ℃ to 37. 5 ℃ | Normal |
| | > 37.5℃ | High |
| Respiration Rate (RR) | <30bpm | Low |
| | 30bpm to 60 bpm | Normal |
| | >60 bpm | High |
| Blood Pressure (DBP) | <90mmHg | Severe |
| | < 120/80 mmHg. | Normal |
| White Blood Cell (DBC) | <8000/µL | Low |
| | <8000/µL to 20,000/ µL | Normal |
| | >20,000/ µL | High |

### 4.3.3 Data Formatting

It is the process of converting the dataset format so that the data mining algorithm can use it or interpret it. The dataset must be structured appropriately for the modeling tool before proceeding with the experimentation. The preprocessed dataset was converted into the dot Attribute Relation File Format (.ARFF), which is appropriate for the data mining tool or algorithm, using WEKA, 3.8.4.

### 4.3.4 Attribute Selection

The majority of data mining algorithms are made to learn from the qualities that are most useful when making decisions. In order to minimize the characteristic space according to a criterion, attribute selection is a procedure that selects a subset of M attributes out of N while adhering to the constraint $M \leq N$. This ensures that the data that are mined are of high quality. Feature selection, another name for attribute selection, is the process of choosing pertinent qualities and eliminating superfluous or unnecessary ones [19]. Therefore, in order to make the process of creating a model simpler, it is need to eliminate such attribute from analysis with the aid of experts and literature reviews.

To rank and select the best features for data mining, the researcher used ‘Gain Ratio Attribute Eval’ Attribute evaluator methods with a ‘Ranker’ search method and the domain expert's advice. The ranked attribute according to the gain ratio feature evaluator method is shown in figure 4.1 below.

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 20 Class):
        Gain Ratio feature evaluator

Ranked attributes:
 0.0261683  12 InVoFr
 0.0169003   9 Dyspnea
 0.0119588   3 Chest Pain
 0.0076717  11 DtoP
 0.0059684   2 sex
 0.0043653   7 Vomitting
 0.0039573  10 SofBreath
 0.0037743  13 AbWBC
 0.0032549  16 Malaise
 0.0029267   6 fever
 0.0020162   4 Chills
 0.0019865  14 Tachypnea
 0.0018137  17 DecBP
 0.0011399   8 LoAppetite
 0.0010436  19 Temp
 0.0007954  15 Rigor
 0.0007567   5 nausea
 0.0000637  18 Resp
 0           1 age
```

**Figure 4.1: Ranked Attributes Based on Gain Ratio Value**

As sown in the figure 4.3, the attributes Increased Vocal Fremitus (InVoFr) score has the maximum gain ratio value and age has the minimum gain ratio value. The domain expert advises the respiration, nausea, and rigor features do not consider major indicators because it also occurs in other pneumonia type of diseases. So, the researcher decides the minimum threshold value of 0.001. And for this research, the researcher approved 16 attributes including the target class out of 19 attributes with the recommendation of domain experts. Hence the selected attributes are Chills, Increased Vocal Fremitus (InVoFr), Dyspnea, chest pain, Dullness to percussion (DtoP), sex, Vomiting, shortness of breath (SoFBreath), white blood cell (AbWBC), malaise, fever, tachypnea, Blood Pressure (DecBP), LoAppetite, Age and Temperature.

Table 4.5: Removing Unnecessary Attributes

| Removed Attributes | |
|---|---|
| Attribute Name | Remark |
| Rigor, Respiration, Nausea | Not Important for The Study |

Attribute selection output

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 16 Class):
        Gain Ratio feature evaluator

Ranked attributes:
 0.02617    1 InVoFr
 0.0169     2 Dyspnea
 0.01196    3 Chest Pain
 0.00767    4 DtoP
 0.00597    5 sex
 0.00437    6 Vomitting
 0.00396    7 SofBreath
 0.00377    8 AbWBC
 0.00325    9 Malaise
 0.00293   10 fever
 0.00202   11 Chills
 0.00199   12 Tachypnea
 0.00181   13 DecBP
 0.00114   14 LoAppetite
 0.00104   15 Temp

Selected attributes: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15 : 15
```

Figure 4.2: Shows Selected Attributes

# CHAPTER FIVE

# EXPERIMENTAL ANALYSIS AND RESULT

## 5.1 Introduction

This chapter conducts a series of experiments and presents the experimentation process in a manner that is simply understood and applied by the organization. This includes describing the actions taken, the decision made, the task completed, the outcome received, and the evaluation of the model findings. To create a classifier model and extract pertinent, useful classification rules, experiments have been conducted. In order to acquire knowledge for the diagnosis of hospital acquired pneumonia (HAP), various experiments were carried out in this study using classification algorithms. As was mentioned in the previous chapter, a total of 3244 data records were pre-processed in order to carry out the experiment

## 5.2 Selecting modeling techniques

The objective of data mining is determining the best model to use. As a result, decision tree, chosen as the classification techniques for model development in order to meet the research objectives. The WEKA 3.8.4 tool environment was used to carry out the analyses. JRip, J48, PART and Random Forest algorithms are of the various possible classification algorithms in WEKA that are used in this study's experiments. The researcher choose the aforementioned algorithms for the following reasons the algorithms have advantages like strong noise tolerance and the capacity to categorize patterns that are not readily apparent, and they are also simple to comprehend and analyse model findings [52].

## 5.3 Experimental setup

One of the tasks in data mining is model development, which is done by providing the processed data to the chosen classification algorithms [53]. The researcher employed two methods to classify the dataset in order to carry out the experiment: percentage split and k-fold (10-folds) cross validation. And four experiments were performed. It involves splitting a given dataset into K pieces or folds, each of which is used as a testing set once. By taking a random sample of scenarios from the learning set without replacement, this partitioning is carried out. The 10-fold

cross-validation method, or K = 10, is used in this investigation. The dataset is therefore divided into ten folds.

The model is trained or developed using the first nine folds in the **first iteration**, and it is tested using the final fold. **The 2$^{nd}$fold** is the testing set and the remaining folds are the training set in the **second iteration**. Until all 10 folds have been used for testing, this process is repeated. The second test option is a percentage split.in this test option, 80 % of the data is used for model development and the remaining 20 % is used for test set. From the total 3244 instances, 2596 instances are used for training set or to develop the model while the remaining 648 instances are used for testing set. The performance of the models is evaluated using the standard metrics of True positive rate, False positive rate, accuracy, precision, receiver operating characteristics (ROC) area, and f measure.

Table 5.1: Experimental Setup

| **Experiments** | **Scenarios** | **Number of attributes** | **Test option** |
|---|---|---|---|
| **Experiment I (Using J48)** | Scenario I | Whole attributes | 10 – fold cross validation |
| | Scenario II | Selected attributes | Percentage split |
| **Experiment II (Using JRip)** | Scenario I | Whole attributes | 10 – fold cross validation |
| | Scenario II | Selected attributes | Percentage split |
| **Experiment III (Using PART)** | Scenario I | Whole attributes | 10 – fold cross validation |
| | Scenario II | Selected attributes | Percentage split |
| **Experiment IV (Using Random Forest)** | Scenario I | Whole attributes | 10 – fold cross validation |
| | Scenario II | Selected attributes | Percentage split |

Table 5.1 presented each experiment's number with their respective scenarios, number of the attribute, and the test options used. Figure 5.1 below shows attributes used for the experiments and the number of instances for each class label.

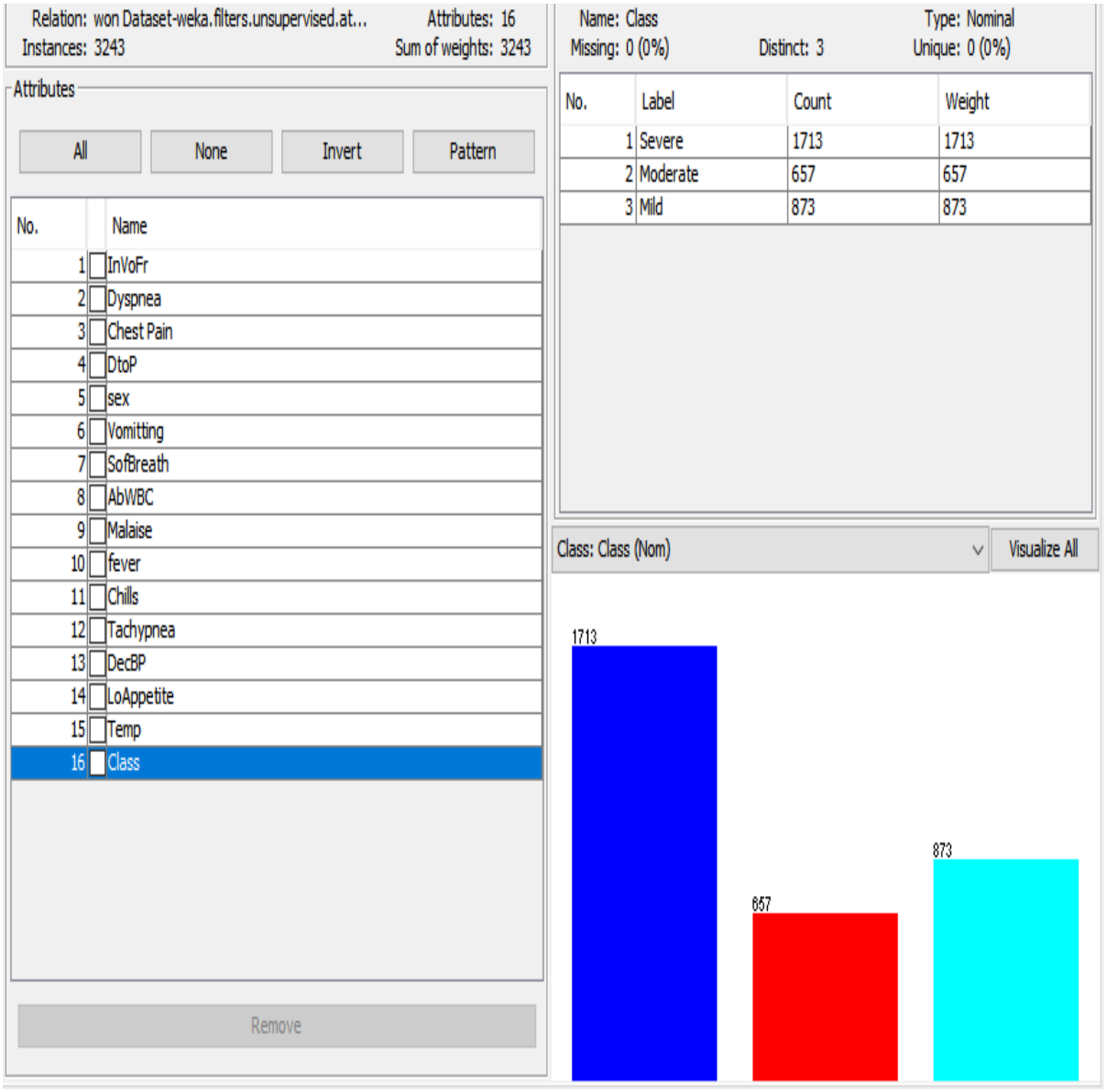



Figure 5.1: Attribute and Number of Instances for the Experiment

## 5.4 Developing classifier model

### 5.4.1 Developing Classifier Model Using J48 Decision tree

On this experimentation two scenarios are conducted. Scenario 1 is performed in whole attributes with 10-fold cross-validation test option. Scenario 2 is performed in selected attributes with percentage split.

**Scenario I**

This experiment is conducted on whole attributes under 10-fold cross validation test option by using the J48 decision tree algorithm. The experimental result show that, from the total 3244 instances, 3132 (96 %) are correctly classified and the remaining 112 (3.453 %) instances are incorrectly classified. And time taken to build the model is 0.3 sec.

Table 5.2: Confusion Matrix for Scenario I

| | Predicted Class | | | |
|---|---|---|---|---|
| | | Severe | Mild | Moderate |
| Actual Class | Severe | 1716 | 18 | 29 |
| | Mild | 23 | 617 | 5 |
| | moderate | 32 | 5 | 798 |

Table 5.2 depicts confusion matrix of the J48 algorithm, the classifier correctly identified 1716 as severe out of 1763 severe instances. And the remaining 18 and 29 identified incorrectly as mild and moderate respectively. The classifier correctly classified 617 as mild out of 645 mild instances and the remaining 23 and 5 instances are identified incorrectly as severe and moderate respectively and the classifier correctly identified 798 as moderate out of 835 moderate instances .and the remaining 32 and 5 were incorrectly classified as severe and mild class respectively. Based on the confusion matrix table 5.3 shows the detailed accuracy of the model.

Table 5.3: Performance Result of Scenario I

| | TPR | FPR | Precision | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| | 0.973 | 0.037 | 0.969 | 0.971 | 0.998 | Severe |
| | 0.957 | 0.009 | 0.964 | 0.960 | 0.999 | Moderate |
| | 0.956 | 0.014 | 0.959 | 0.957 | 0.999 | Mild |
| Weighted Avg. | 0.965 | 0.026 | 0.965 | 0.965 | 0.998 | |

The weighted average of True Positive Rate indicates the average performance of the model in classifying Hospital Acquired Pneumonia disease example severe as severe, moderate as moderate and mild as mild. The weighted average of False Positive Rate indicates an error rate of the classifier can classify incorrectly as severe Moderate and Mild. The weighted average precision metric shows the accuracy of the positive class. This shows how many of the correctly predicted cases actually turned out to be positive. The weighted average of the ROC Area indicates the tradeoff between weighted average True Positive Rate and Weighted average False Positive Rate.

**Scenario II**

This experimentation is carried out on selected attributes with percentage split test option by using the J48 decision tree Algorithm. The number of instances correctly classified by the J48 algorithm is 643 (99.07%) and incorrectly classified Instances are 6 (0.9%) from a total number of 649 of testing instances. Time taken to test the model is 0.48 sec and the confusion matrix of the model is shown below table 5.4 and the detailed accuracy of the developed model is presented in table 5.5 below.

Table 5.4: Confusion Matrix for Scenario II

| | Predicted Class | | | |
|---|---|---|---|---|
| | | Severe | Mild | Moderate |
| Actual Class | Severe | 339 | 1 | 2 |
| | Mild | 1 | 124 | 2 |
| | moderate | 1 | 1 | 791 |

Table 5.5 depicts confusion matrix of J48 algorithm, the classifier correctly classified 339 as severe out of 342 severe instances. And the remaining 1 and 2 identified incorrectly as mild and moderate respectively. The classifier correctly classified 124as mild out of 127 mild instances and the remaining 1and 2 instances are identified incorrectly as severe and moderate respectively and the classifier correctly identified 791 as moderate out of 793 moderate instances and the remaining 1 and 1were incorrectly identified as severe and mild respectively. Based on the confusion matrix table 5.5 shows the detailed accuracy of the model

Table 5.5: Performance Result of Scenario II

| | TPR | FPR | Precision | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| | 0.824 | 0.220 | 0.805 | 0.814 | 0.871 | Severe |
| | 0.673 | 0.080 | 0.632 | 0.652 | 0.888 | Moderate |
| | 0.714 | 0.093 | 0.772 | 0.742 | 0.894 | Mild |
| Weighted Avg. | 0.764 | 0.157 | 0.765 | 0.764 | 0.881 | |

### 5.4.2 Developing Classifier Model Using JRip Decision tree

On this experimentation two scenarios are conducted. Scenario 1 is performed in whole attributes with 10-fold cross-validation test option. Scenario 2 is performed in selected attributes with percentage split.

**Scenario I**

This experiment is conducted on whole attributes under 10-fold cross validation test option by using the JRip decision tree algorithm. The experimental result show that, from the total 3244 instances, 2945 (90.8 %) are correctly classified and the remaining 298 (9.1 %) instances are incorrectly classified. And time taken to build the model is 0.06 sec. the confusion matrix of the model is shown below table 5.6 and the detailed accuracy of the developed model is presented in table 5.7 below.

Table 5.6: Confusion Matrix for Scenario I

| Actual Class | Predicted Class | | | |
|---|---|---|---|---|
| | | Severe | Mild | Moderate |
| | Severe | 1687 | 24 | 52 |
| | Mild | 88 | 540 | 17 |
| | moderate | 108 | 9 | 718 |

Table 5.6 depicts confusion matrix of J48 algorithm, the classifier correctly classified 1687 as severe out of 1763 severe instances. The remaining 24 and 52 identified incorrectly as mild and moderate respectively. The classifier correctly classified 540 as mild out of 645 mild instances and the remaining 88 and 17 instances are identified incorrectly as severe and moderate respectively and the classifier correctly identified 718 as moderate out of 835 moderate instances and the remaining 108 and 9 were incorrectly identified as severe and mild respectively. Based on the confusion matrix table 5.6 below shows the detailed accuracy of the model.

Table 5.7: Performance Result of Scenario I

| | TPR | FPR | Precision | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| | 0.957 | 0.132 | 0.896 | 0.925 | 0.975 | Severe |
| | 0.837 | 0.013 | 0.942 | 0.887 | 0.989 | Moderate |
| | 0.860 | 0.029 | 0.912 | 0.885 | 0.980 | Mild |
| Weighted Avg. | 0.908 | 0.082 | 0.909 | 0.907 | 0.979 | |

**Scenario II**

Using the JRip classification method, this experimentation is carried out on a subset of selected attributes with a percentage split test option. The JRip classifier algorithm accurately classifies 607(93.5%) out of the 649 testing instances, while 42(6.4%) instances are incorrectly classified. And time taken by the algorithm to build the model is 0.02 sec.the confusion matrix of the model is shown below table 5.8 and the detailed accuracy of the developed model is presented in table 5.9 below.

Table 5.8: Confusion Matrix for Scenario II

| | Predicted Class | | | |
|---|---|---|---|---|
| | | Severe | Mild | Moderate |
| Actual Class | Severe | 304 | 12 | 24 |
| | Mild | 88 | 14 | 8 |
| | moderate | 140 | 6 | 53 |

Table 5.8 shows confusion matrix of JRip algorithm, the classifier correctly classified 304 instances as severe out of 340 severe instances. And the remaining 12 instances were incorrectly identified as mild and 24 as moderate. The classifier correctly classified 14 as mild out of 110 mild instances and the remaining 88 and 8 instances are identified incorrectly as severe and moderate respectively and the classifier correctly classified 53 instances as moderate out of 199 moderate instances and the remaining 140 and 6 were incorrectly identified as severe and mild respectively. Based on the confusion matrix table 5.8 below shows the detailed performance result of the model.

Table 5.9: Performance Result of Scenario II

| | TPR | FPR | Precision | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| | 0.894 | 0.738 | 0.571 | 0.697 | 0.635 | Severe |
| | 0.127 | 0.033 | 0.438 | 0.197 | 0.620 | Moderate |
| | 0.266 | 0.071 | 0.624 | 0.373 | 0.680 | Mild |
| Weighted Avg. | 0.572 | 0.414 | 0.565 | 0.513 | 0.647 | |

### 5.4.3 Developing Classifier Model Using PART Decision tree

On this experimentation three scenarios are conducted. Scenario 1 is performed in whole attributes with 10-fold cross-validation test option. Scenario 2 is performed in selected attributes with percentage split.

## Scenario I

This experiment is conducted on whole attributes under 10-fold cross validation test option by using PART decision tree algorithm. The experimental result show that, from the total 3244 instances, 3131(97%) are correctly classified and the remaining 113 (3.4 %) instances are

incorrectly classified. And time taken to build the model is 0.48 sec. the confusion matrix of the model is shown below table 5.10 and the detailed accuracy of the developed model is presented in table 5.11 below.

Table 5.10: Confusion Matrix for Scenario I

| | Predicted Class | | | |
|---|---|---|---|---|
| | | Severe | Mild | Moderate |
| Actual Class | Severe | 1726 | 13 | 24 |
| | Mild | 28 | 613 | 4 |
| | moderate | 38 | 6 | 791 |

Table 5.9 shows confusion matrix of PART algorithm, the classifier correctly classified 1726 instances as severe out of 1763 severe instances. And the remaining 13 instances were incorrectly identified as mild and 24 as moderate. The classifier correctly classified 613 as mild out of 645 mild instances and the remaining 28 and4 instances are classified incorrectly as severe and moderate respectively and the classifier correctly classified 791 instances as moderate out of 835 moderate instances and the remaining 38 and 6 were incorrectly classified as severe and mild respectively. Based on the confusion matrix table 5.9 below shows the detailed performance result of the model.

Table 5.11: Performance Result of Scenario I

| | TPR | FPR | Precision | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| | 0.979 | 0.045 | 0.963 | 0.971 | 0.998 | Severe |
| | 0.950 | 0.007 | 0.970 | 0.960 | 0.999 | Moderate |
| | 0.947 | 0.012 | 0.966 | 0.956 | 0.999 | Mild |
| Weighted Avg. | 0.965 | 0.029 | 0.965 | 0.965 | 0.998 | |

**Scenario II**

Using the PART classification method, this experimentation is carried out on a subset of attributes with a percentage split test option. The PART classifier algorithm accurately classifies 644(99.3%) out of the 649 testing instances, while 5(0.8%) instances are incorrectly classified. And time taken by the algorithm to build the model is 0.04 sec. the confusion matrix of the model is shown below table 5.12 and the detailed accuracy of the developed model is presented in table 5.13 below.

Table 5.12: Confusion Matrix for Scenario II

| | Predicted Class | | | |
|---|---|---|---|---|
| | | Severe | Mild | Moderate |
| Actual Class | Severe | 339 | 1 | 2 |
| | Mild | 1 | 126 | 1 |
| | moderate | 1 | 1 | 179 |

Table 5.11 shows confusion matrix of PART algorithm, the classifier correctly classified 339 instances as severe out of 342 severe instances. And the remaining 1 instance incorrectly identified as mild and 2 as moderate. The classifier correctly classified 126 as mild out of 128 mild instances and the remaining 1 and 1 instance are classified incorrectly as severe and moderate respectively and the classifier correctly classified 179 instances as moderate out of 181 moderate instances and the remaining 1 and 1 were incorrectly classified as severe and mild respectively. Based on the confusion matrix table 5.11 below shows the detailed performance result of the model.

Table 5.13: Performance Result of Scenario I

| | TPR | FPR | Precision | F-Measure | ROC area | Class |
|---|---|---|---|---|---|---|
| | 0.991 | 0.001 | 1.000 | 0.996 | 1.000 | Severe |
| | 0.992 | 0.004 | 0.984 | 0.988 | 1.00 | Moderate |
| | 0.994 | 0.006 | 0,984 | 0.989 | 1.00 | Mild |
| **Weighted Avg** | 0.992 | 0.003 | 0.992 | 0.992 | 1.00 | |

### 5.4.4 Developing Classifier Model Using Random Forest Tree

On this experimentation two scenarios are conducted. Scenario 1 is performed in whole attributes with 10-fold cross-validation test option. Scenario 2 is performed in selected attributes with percentage split**.**

**Scenario I**

This experiment is conducted on whole attributes under 10-fold cross validation test option by using Random Forest decision tree algorithm. The experimental result show that, from the total 3244 instances, 3103(96%) are correctly classified and the remaining 140 (4.3 %) instances are incorrectly classified. And time taken by the algorithm to build the model is 0.12 sec. the confusion matrix of the model is shown below table 5.14 and the detailed accuracy of the developed model is presented in table 5.15 below.

Table 5.14: Confusion Matrix for Scenario II

| | Predicted Class | | | |
|---|---|---|---|---|
| | | Severe | Mild | Moderate |
| Actual Class | Severe | 1669 | 15 | 29 |
| | Mild | 31 | 621 | 5 |
| | moderate | 50 | 10 | 813 |

Table 5.14 shows confusion matrix of random forest algorithm, the classifier correctly classified 1669 instances as severe out of 1713 severe instances. And the remaining 15 instances were incorrectly identified as mild and 29 as moderate. The classifier correctly classified 621 as mild out of 657 mild instances and the remaining 31 and 5 instances are classified incorrectly as severe and moderate respectively and the classifier correctly classified 813 instances as moderate out of 873 moderate instances and the remaining 50 and 10 were incorrectly classified as severe and mild respectively. Based on the confusion matrix table 5.15 below shows the detailed performance result of the model.

Table 5.15: Performance Result of Scenario I

| | TPR | FPR | Precision | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| | 0.974 | 0.053 | 0.954 | 0.964 | 0.993 | Severe |
| | 0.945 | 0.010 | 0.961 | 0.953 | 0.997 | Moderate |
| | 0.931 | 0.014 | 0.960 | 0.945 | 0.996 | Mild |
| Weighted Avg. | 0.957 | 0.034 | 0.957 | 0.957 | 0.995 | |

**Scenario II**

Using the random forest classification method, this experimentation is carried out on a subset of attributes with a percentage split test option. The random forest classifier algorithm accurately classifies 627(97%) out of the 649 testing instances, while 22(3%) instances are incorrectly classified. And time taken by the algorithm to build the model is 0.02 sec. The confusion matrix of the model is shown below table 5.16 and the detailed accuracy of the developed model is presented in table 5.17 below.

Table 5.16: Confusion Matrix for Scenario II

| | Predicted Class | | | |
|---|---|---|---|---|
| | | Severe | Mild | Moderate |
| Actual Class | Severe | 339 | 1 | 3 |
| | Mild | 8 | 118 | 1 |
| | moderate | 7 | 3 | 170 |

Table 5.12 shows confusion matrix of PART algorithm, the classifier correctly classified 339 instances as severe out of 343 severe instances. And the remaining 1 instance incorrectly identified as mild and 3 as moderate. The classifier correctly classified 118 as mild out of 127 mild instances and the remaining 8 and 1 instance are classified incorrectly as severe and moderate respectively and the classifier correctly classified 170 instances as moderate out of 180 moderate instances and the remaining 7 and 3 were incorrectly classified as severe and mild respectively. Based on the confusion matrix table 5.17 below shows the detailed performance result of the model.

Table 5.17: Performance Result of Scenario I

| | TPR | FPR | Precision | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| | 0.991 | 0.049 | 1.958 | 0.974 | 0.997 | Severe |
| | 0.945 | 0.006 | 0.975 | 0.952 | 0.997 | Moderate |
| | 0.944 | 0.009 | 0.977 | 0.960 | 0.997 | Mild |
| Weighted Avg. | 0.966 | 0.029 | 0.966 | 0.966 | 0.997 | |

## 5.5 Performance Comparison of Classifier Model

One of the objectives of this study is to select the best classifier model after a model has been built, one that can identify the Hospital Acquired Pneumonias situation as severe, moderate, or mild. Using the 10-fold cross-validation test choices and percentage split utilized for conducting tests in the WEKA data mining tool, J48, JRip, Random Forest and PART algorithms were employed in order to pick the best classifier model. Table 5.18 the following lists the four chosen classification algorithms along with their corresponding best performance accuracy.

Table 5.18: Performance Comparison of the Classified Model

| Algorithm Used | Correctly Classified Instances | | Incorrectly Classified Instances | | Time Taken (sec) |
|---|---|---|---|---|---|
| **10-fold Cross Validation Test Option** | | | | | |
| | No | Accuracy | No | Accuracy | |
| J48(Scenario I) | 3132 | 96% | 112 | 3.4% | 0.3 |
| JRip (Scenario I) | 2945 | 90.8% | 298 | 9.1% | 0.6 |
| PART (Scenario I) | 3131 | 97% | 113 | 3.4% | 0.48 |
| Random Forest (Scenario I) | 3103 | 96% | 140 | 4.3% | 0.12 |
| **Percentage Split Test Option** | | | | | |
| | No | Accuracy | No | Accuracy | |
| J48(Scenario II) | 643 | 99.07% | 6 | 0.9% | 0.4 |
| JRip (Scenario II) | 607 | 94% | 42 | 6.4% | 0.02 |
| PART (Scenario II) | 644 | **99.3%** | 5 | 0.8% | 0.04 |
| Random Forest (Scenario II) | 627 | 97% | 22 | 3% | 0.02 |

In these experiment four algorithms are used J48, JRip, PART and Random Forest decision tree-based classifiers. From each method a total of eight models are developed based on cross validation and percentage split test option. As discussed in chapter two ROC Areas, Accuracy, True Positive Rate, False Positive Rate, Precision, and other evaluation metrics were used to compare and evaluate the model's overall performance based on individual outcomes. Table 5.18 above shows the result of the model.

From **experiment I** using the J48 decision tree algorithm scenario two has performed well, which was conducted on selected attributes with percentage split. From **experiment II** using JRip rule induction, the second scenario has performed well, which was conducted on selected attributes with percentage split. From **experiment III** using the PART rule induction algorithm, the second scenario has performed well, which was conducted on selected attributes with percentage split. From **experiment IV** using random forest rule induction scenario two has performed well, which was conducted on selected attributes with percentage split.

Thus, based on this result and assessment metrics employed in this study, the model created by the PART classifier algorithm conducted on selected attributes with percentage split test option was determined to be the best classifier model.so the researcher used this classifier for extracting rules.

## 5.6 Rule Extraction from PART Rule Induction

Among many methods, the partial decision tree (PART) classifier was found to be the most effective model with chosen attributes under selected attributes with percentage split. Constructing the knowledge-based system comes after the partial decision tree classification algorithm has produced the rules. To select the most effective rules that cover the majority of the dataset in the domains, the researcher discussed with domain experts. The knowledge obtained from the domain experts is combined with the chosen rules to create knowledge-based systems.

**Rule 1:** IF fever =No AND Malaise= Yes AND Temperature=High AND InVoFr=Yes AND Chest Pain =Stabbing AND Vomiting =yes: THEN Severe (12.0/1.0)

**Rule 2:** IF Fever =No AND Tachypnea= Yes AND DtoP =Dull: THEN Mild (6.0/1.0)

**Rule 3:** IF Fever = No AND Tachypnea=Yes AND Temperature =Low AND Malaise =Yes AND SoFBreath=No AND DtoP Hyper rmonant AND Chest Pain= Stabbing: THEN Severe (8.0)

**Rule 4:** If InVoFr =Yes AND Fever =Yes AND Malaise =No AND Sex =Male AND DecBP= Diastolic AND Temperature =Normal AND AbWBC= Leukopenia: THEN Moderate (3.0)

**Rule 5:** If InVoFr =Yes AND Fever =Yes AND Malaise =No AND DecBP= Diastolic AND Vomiting =No AND Chest Pain = Stabbing: THEN Severe (6.0/1.0)

**Rule 6:** If InVoFr =Yes AND Fever =Yes AND Malaise =Yes AND DtoP =Dull AND Temperature =Normal AND =LoAppetite = Yes: THEN Moderate (3.0/1.0)

**Rule 7:** If InVoFr =Yes AND Fever =Yes AND Malaise =Yes AND DtoP = Hyper rmonant AND Sex=M Chest Pain =Sharp AND Tachypnea = Yes: THEN Severe (14.0/3.0)

**Rule 8:** If Malaise= Yes AND Tachypnea = Yes AND Dyspnea =No AND DtoP = Hyper rmonant AND AbWBC =Leukopenia: THEN Mild (7.0)

**Rule 9:** If InVoFr= Yes AND Malaise = Yes AND Tachypnea =Yes AND AbWBC =Leukocytosis AND Temperature =Low AND Sex=F AND Dyspnea =Yes: THEN Mild (5.0)

**Rule 10:** If InVoFr= Yes AND Malaise = Yes AND Tachypnea =Yes AND AbWBC =Leukocytosis AND Dyspnea =No AND Fever=No AND DtoP =Normal AND LoAppetite =Yes: THEN Mild (7.0/2.0)

These are sample rules which are generated under PART rule induction. The rules indicate that **Rule1**: if a patient has fever and there is symptom of malaise and if temperature is high and increased vocal fremitus is yes and if Chest Pain is Stabbing AND Vomiting is yes then the patient is diagnosed as Severe (12.0/1.0). according to the rule out of 12 instances 11 instances are correctly classified and its accuracy to classify as severe is 96% and 1 instance is incorrectly classified. The Rule indicates that **Rule7**: if a patients increased vocal fremitus is yes and fever is yes and if malaise yes and Dullness to percussion is Hyper rmonant and if chest pain is sharp and tachypnea is yes then the patient is diagnosed as severe. According to the rule out of 14 instances 11 are correctly classified and the remaining 3 are incorrectly classified and its accuracy to classify as severe is 98%.

## 5.7 Knowledge Extraction from Expert

Knowledge acquisition and reasoning are crucial components of intelligent systems in the field of artificial intelligence, particularly knowledge base systems and expert systems. Acquiring knowledge is a challenging task because multiple approaches and strategies depending on the expert domain, knowledge type and knowledge engineer are needed to transmit expert information [44]. In order to generate the KBS for this study, knowledge was extracted both manually (by domain expert interviews) and automatically (through data mining). Thus, in order to gather information, non-structured interview type is followed with an expert because it is used to acquire a preliminary problem-solving knowledge and the interviewer raised many questions concerning the focus area as described below.

What is Hospital Acquired Pneumonia? "Hospital-acquired (nonsocial) pneumonia (HAP) is a pneumonia that occurs 48 hours or more after admission, Alveoli, which are tiny air sacs in the lungs; fill with air when a healthy individual breathes and has different severity levels" So what are the different severity levels of Hospital Acquired Pneumonia? "This disease has different severity levels and classified and assessed based on the symptoms that we have observed and lab result namely severe, mild and moderate". Which severity level of the diseases is more commonly happened in this hospital? "All are frequently happened but comparatively the most common one is severe hospital acquired pneumonia". What are the different symptoms that make it differ from other pneumonia type "It has different symptoms that differ from other pneumonia type such as tachypnea, dyspnea, loss of appetite, shortness of breath". What are the symptoms or main signs for diagnosis and treatment recommendation of Hospital Acquired Pneumonia? "There are different signs are there but the most commons are Chills, Fever, Malaises, Loss of Appetite, Vomiting, Chest Pain, Shortness of Breath, Decreased Blood Pressure, Dyspnea, Tachypnea, Dullness to Percussion and an Increased Vocal Fremitus".

What type of treatment recommendation is given for each severity level? After the diseases is diagnosed treatment recommendation is given this treatment is depend on the severity level of the diseases and age of the patient for **severe** level Piperacillin /tazobactam 4.5 g IV 8 hourly plus Gentamicin 5 mg/kg/day IV and for pregnant woman Amoxicillin 500 mg PO TID for 07

days and Amoxicillin 15-30 mg/kg TID for 07days for child is given. And if the diseases are diagnosed as **moderate** Cephazolin 1 g IV 8hourly plus Gentamicin 5 mg/kg/day IV and for pregnant woman Cephalexin 500 mg PO BID for 07days and Azithromycin 30mg/kg once daily for 05days for child is given. If it is **mild** level of hospital acquired pneumonia then amoxycillin + clavulanic acid 875/125 mg (1 tablet) orally for 12 hourly recommended. And for pregnant woman Amoxicillin 500 mg PO TID for 07 days. Cephalexin 500 mg PO BID for 07days Clindamycin 300mg PO QID for 10days are recommended whereas for child's Amoxicillin 15-30 mg/kg TID for 07days Azithromycin 30mg/kg once daily for 05days are recommended.

## 5.8 Expert Knowledge modeling

Models break down complex systems into simpler, more manageable components that are simple to comprehend and work with in order to capture their important characteristics. Most knowledge is unstructured and frequently in the form of tacit knowledge when it is first acquired. The knowledge engineer will make an effort to comprehend both explicit and tacit information, and will then utilize straightforward visual aids to encourage dialogue among subject matter experts.

As described in chapter three the researcher applied the decision tree knowledge method for modelling in this investigation. Figure5.2 below shows the knowledge acquired from the domain expert in the form of decision tree for the diagnosis of hospital acquired pneumonia.

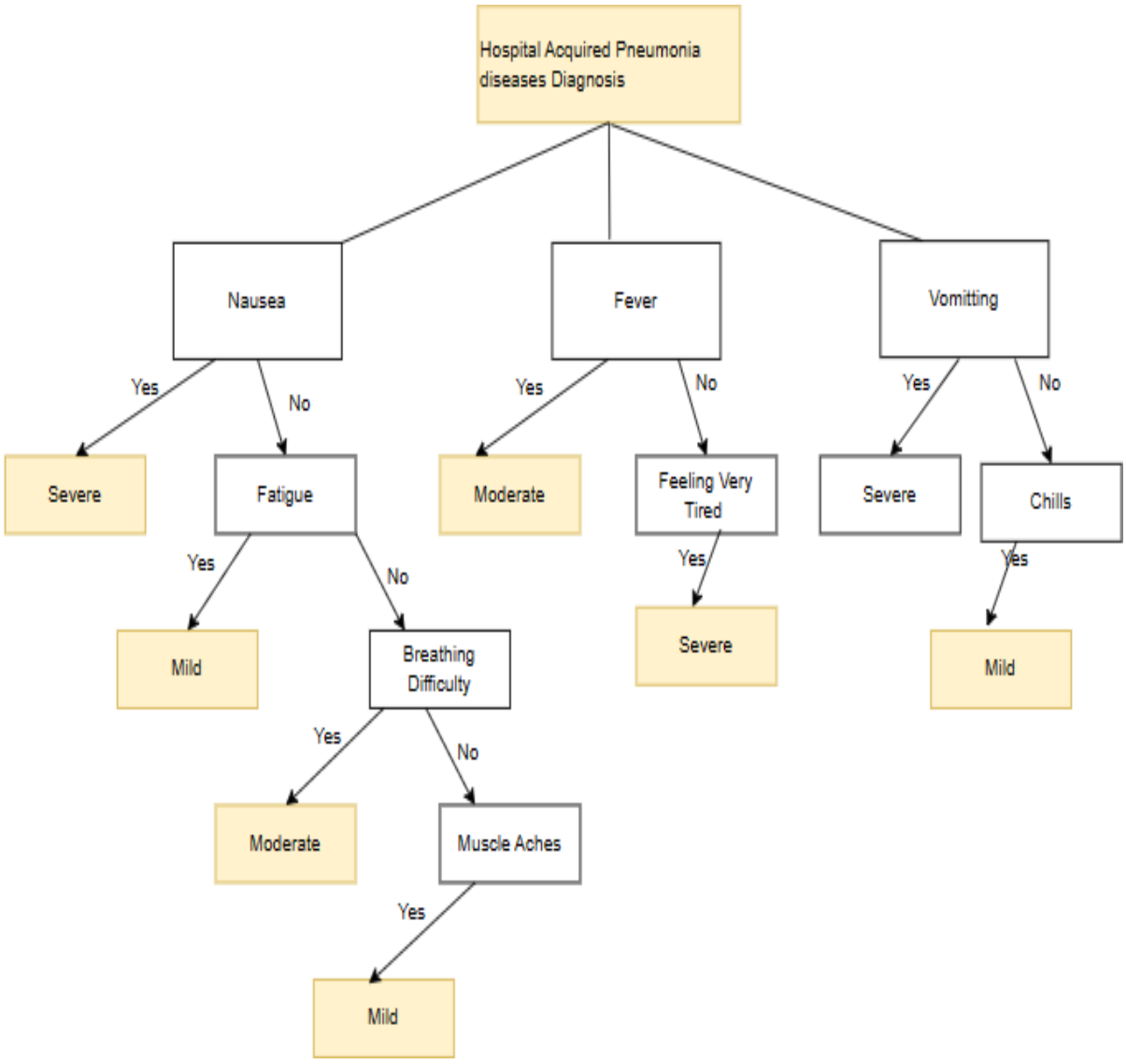



Figure 5.2: Decision Tree for HAP diagnosis Acquired from Domain Expert

After diagnosis of the diseases the next phase is providing an appropriate treatment recommendation of Hospital Acquired Pneumonia (HAP) diseases. Treatment recommendation of the diseases includes the following medications: -amoxycillin + clavulanic acid 875/125 mg (1 tablet) orally 12 hourly, Cephazolin 1 g IV 8 hourly plus Gentamicin* 5 mg/kg/day IV and Piperacillin /tazobactam 4.5 g IV 8 hourly plus Gentamicin5 mg/kg/day IV figure 5.5 below shows the type of treatment that are recommended for each severity level of the diseases.

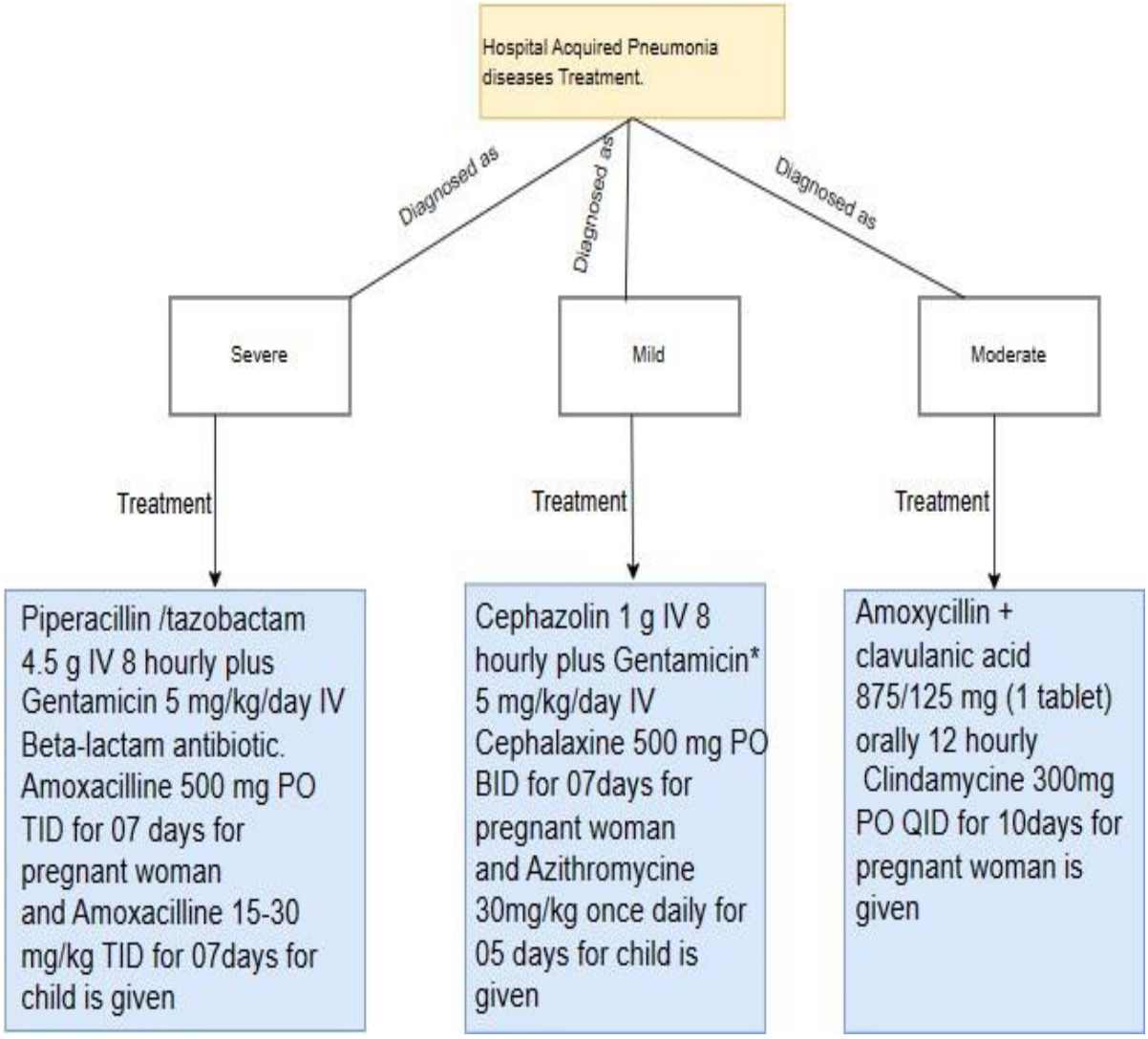



Figure 5.3: Decision Tree for HAP treatment acquired from Domain Expert

## 5.9 Expert Knowledge Representation

In this study, the knowledge obtained from domain experts was represented in the **IF-THEN** format. This is because after it has been gathered the knowledge is represented using the rule-based knowledge representation method. In general, the following arrangement was used to illustrate Figures 5.4 and 5.5 above.

**Rules extracted from expert for the diagnosis of Hospital Acquired Pneumonia**

**Rule 1:** IF patient has Nausea, THEN the diseases =Severe HAP.

**Rule 2:** IF patient has No Nausea AND Muscle Fatigue Yes THEN the diseases =Mild HAP.

**Rule 3:** IF patient has No Nausea with muscle fatigue No and breathing difficulty is Yes THEN the diseases =Moderate HAP.

**Rule 4:** IF patient has Fever, THEN the diseases =Moderate HAP.

**Rule 5:** IF patient has No fever with Feeling very tiered is Yes, THEN the disease =Severe HAP.

**Rule 6:** IF patient has vomiting THEN the diseases =Severe HAP.

**Rule 7:** IF patient has No vomiting AND chills Yes THEN the diseases =Mild HAP.

**Rules extracted from expert for the treatment of Hospital Acquired Pneumonia**

**Rule 1:** IF the diagnosis results of HAP =Severe, THEN the treatment= Piperacillin /tazobactam 4.5 g IV 8 hourly plus Gentamicin5 mg/kg/day IV, Beta-Lactam antibiotic, Piperacillin. And Amoxicillin 500 mg PO TID for 07 days for pregnant woman and Amoxicillin 15-30 mg/kg TID for 07days for children is given.

**Rule 2:** IF the diagnosis results of HAP =Mild, THEN the treatment= Cephazolin 1 g IV 8 hourly plus Gentamicin* 5 mg/kg/day IV. And Cephalexin 500 mg PO BID for 07days for pregnant woman and Azithromycin 30mg/kg once daily for 05days for children is given.

**Rule 3:** IF the diagnosis results of HAP =Moderate, THEN the treatment=amoxycillin + clavulanic acid 875/125 mg (1 tablet) orally 12 hourly and Clindamycin 300mg PO QID for 10days is given for pregnant woman.

# CHAPTER SIX

# IMPLEMENTATION AND DISCUSSION OF RESULT

This chapter provides a brief overview of knowledge-based system building, system evaluation, and findings discussion. Following the acquisition of the required knowledge from two sources (experts and data mining), it is expressed by IF...THEN rules. Following representation, the two source knowledge sets are combined, and the represented knowledge is implemented using the Prolog programming language in a way that makes sense to the inference engine. Java programming is used to create the user interface that allows users to interact with the knowledge base. Java Net Beans is used for the GUI and SWI-prolog is used for knowledge-based system development[54].

## 6.1 Implementation of discovered rule to KBS

Using the expertise of subject experts and a collaborative predictive model, this work built the KBS for the diagnosis and treatment advice of Hospital Acquired Pneumonia (HAP) disease. Thus, creating or developing the KBS comes next once the modeling, representation, and knowledge collection processes are finished. Forty-five and seven rules, respectively, are produced by PART classification algorithms and domain specialists to diagnose neonatal disorders as severe, mild, and moderate. Additionally, the knowledge obtained from domain specialists was useful for addressing each condition. The majority of the rules in the knowledge base employed many variables, although others just used one variable with its corresponding values.

### 6.1.1 Structure of PART classifiers and Prolog

PART rule induction classifier was selected as the best algorithm from the four classification algorithms tested in the experimentation.it generates rule in the form of IF –then format. If (Condition) …then (Conclusion). two or more conditions are joined by “AND” and after the “condition”: “meaning implies follows. Then the conclusion part indicates the class of the diseases [55].

Table 6.1: Sample Rule Generated by PART Algorithm

| Rule No | Rules | |
|---|---|---|
| | Conditions | Conclusion |
| 1 | fever =No AND Malaise= Yes AND Temperature=High AND InVoFr=Yes AND Chest Pain =Stabbing AND Vomiting =yes: | Severe (12.0/1.0) |
| 2 | InVoFr =Yes AND Fever =Yes AND Malaise =No AND Sex =Male AND DecBP= Diastolic AND Temperature =Normal AND AbWBC= Leukopenia: | Moderate (3.0) |
| 3 | Malaise= Yes AND Tachypnea = Yes AND Dyspnea =No AND DtoP = Hyper rmonant AND AbWBC =Leukopenia: | Mild (7.0) |

The above rule 6.1 can be interpreted as follows:

**Rule 1:** IF fever =No AND Malaise= Yes AND Temperature=High AND InVoFr=Yes AND Chest Pain =Stabbing AND Vomiting =yes: THEN it is Severe (12.0/1.0)

If those conditions are true means if there is malaise, Increased Vocal fremitus and vomiting with high temperature and stabbing chest pain and no fever, then the conclusion is Severe HAP. But if one of the conditions is false then the conclusion is false.

**Rule 2:** IF InVoFr =Yes AND Fever =Yes AND Malaise =No AND Sex =Male AND DecBP= Diastolic AND Temperature =Normal AND AbWBC= Leukopenia: THEN it is Moderate (3.0).

**Rule 3:** IF Malaise= Yes AND Tachypnea = Yes AND Dyspnea =No AND DtoP = Hyper rmonant AND AbWBC =Leukopenia: THEN it is Mild (7.0).

But prolog doesn't work in IF…THEN condition but rather in reverse order that means it start with conclusion and then moves on to the facts that can be used to verify the goal is correct. So, the rule is reversed from IF-THEN to THEN-IF to implement in prolog.

**Therefore**, the above rule has to be formatted as:

**Severe:** -Fever =No, Malaise =Yes, Temperature =High, InVoFr=Yes, Chest Pain =Stabbing, Vomiting =Yes.

**Moderate:** -InVoFr =Yes, Fever =Yes, Malaise =No, Sex =Male, DecBP= Diastolic, Temperature =Normal, AbWBC= Leukopenia.

**Mild:** -Malaise= Yes, Tachypnea = Yes, Dyspnea =No, DtoP = Hyper rmonant, AbWBC =Leukopenia.

A Prolog rule has the form**:** Head**: -** Body. The body of each rule is a prolog goal. A goal is a Prolog term that denotes a predicate and its arguments. The conclusion comes first followed by "," replacing "AND" in the PART rule. Then, prolog rules terminate by period (.)

## 6.2 KBS Development

The proposed framework figure 3.3 in chapter three shows that, after the knowledge is acquired from data mining and domain expert the next task is building Knowledge Base System for diagnosis and treatment recommendation of Hospital Acquired Pneumonia (HAP). Then the obtained Knowledge is programmed in the Knowledge base as rules about the subject and knowledge relationship in terms of if-then rules. The developed KBS has four main parts these are, Knowledge base, explanation facility, inference engine and user interface [32].

### 6.2.1 Knowledge base

The set of rules, facts, and relationship or the encoded knowledge about diagnosis and treatment recommendation of selected Hospital Acquired Pneumonia (HAP). After the knowledge is represented in the form of rules-based representation techniques, the rules are codified to the knowledge base of the base of the prototypes system using the prolog programming language.

### 6.2.2 Inference Engine

It involves approaches for problem-solving and it is used to analyze the rules and knowledge contained in the knowledge base that is used to provide a logical conclusion.it is the backbone of Knowledge base System because it consists of the methods that the system solves the problem by using this techniques.it deduces facts and give conclusion for the knowledge base based on the user input and facts from the knowledge. The inference engine is built based on backward chaining mechanism because the inference engine of a prolog programming language is built-in with backward-chaining inference strategy.

### 6.2.3 User Interface

This is the interaction between the users and Knowledge-based system. The knowledge-based system interface is built by java Net Beans. The user asks the system a series of questions then by interacting through the interface, the system responds the request. Figure 6.1 below shows that the user interfaces for diagnosis and treatment recommendation of Hospital Acquired Pneumonia (HAP) [56].

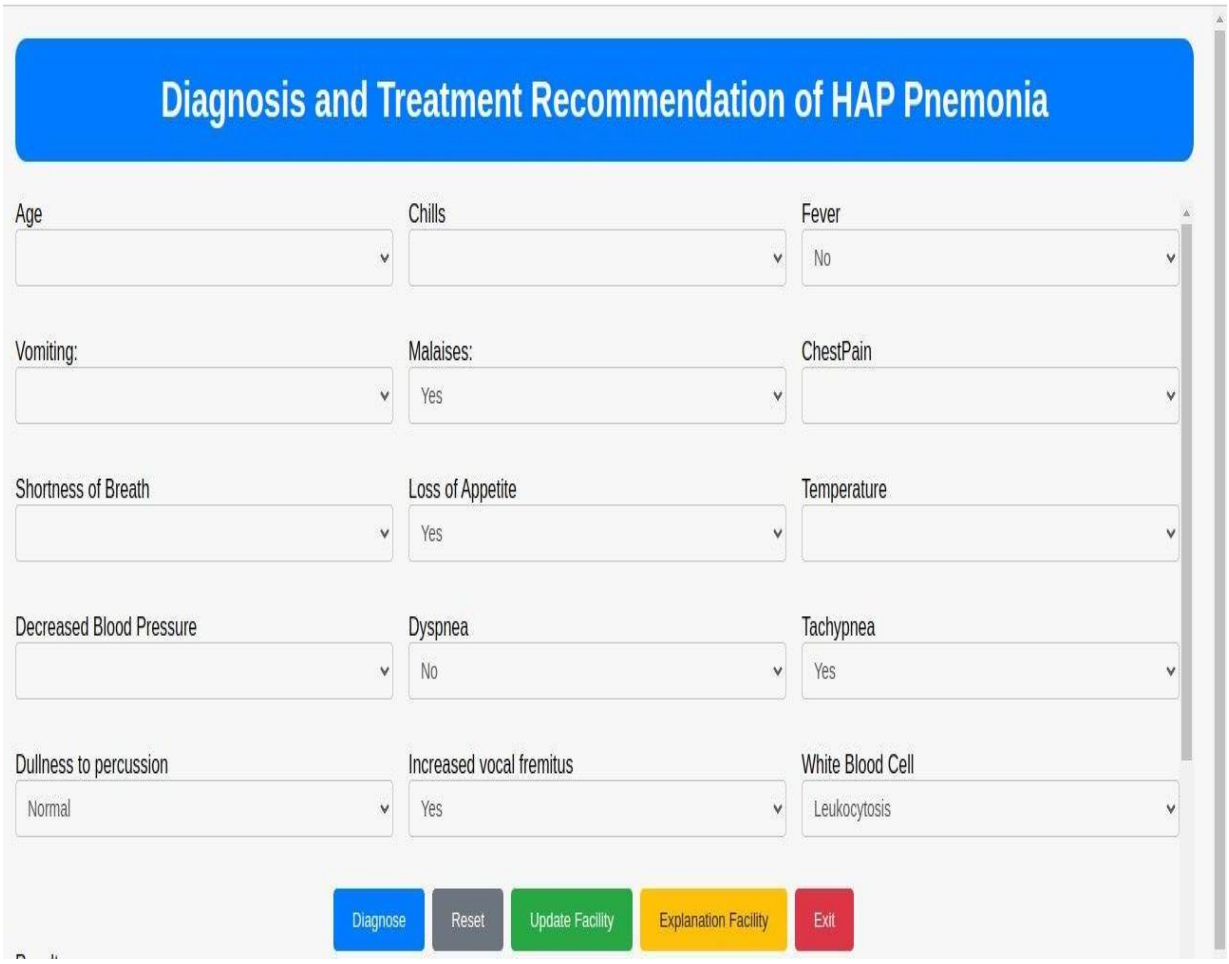



Figure 6.1: GUI of the Developed KBS

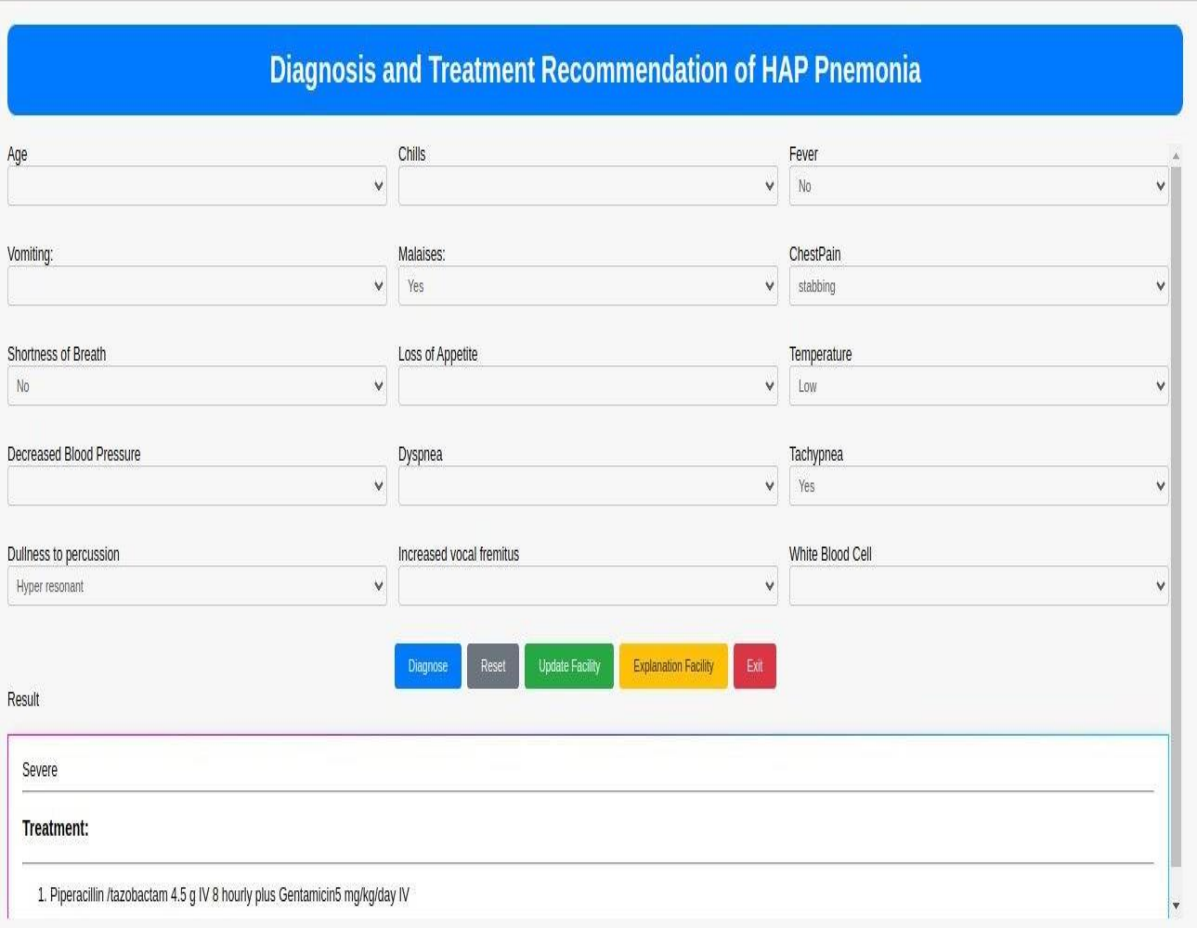



Figure 6.2: GUI of Treatment Recommendation

## 6.3 Learning System

The development of useful techniques that allow users add new information and/or rules to the system is a crucial topic for knowledge-based systems. The created system may pick up new information about symptoms and indicators without the need for a knowledge engineer to modify the code. The domain experts in this study can quickly add new rules to the knowledge base that are utilized for diagnosing hospital acquired pneumonia (HAP) and recommending therapies. Any undiscovered rules and treatments, or newly discovered treatments, are uploaded to the knowledge base so that the Inference system can use the new rules to draw conclusions. Experts can update the facility by selecting the update option[57].

## 6.4 Evaluation of the System

The next step in this work is to test and evaluate the knowledge-based system prototype for the diagnosis and treatment recommendations of Hospital Acquired Pneumonia (HAP), which was developed earlier. Because the knowledge engineer uses the evaluation of the knowledge-based system as the last phase to determine whether or not the prototype system is accurate and useful to users. The evaluation is carried out to verify the system's ease of use and acceptance by users as well as to determine whether the goal of the research project has been met. We can test and evaluate a knowledge-based system as long as we know what to expect, since the goal of the prototype system's testing and evaluation is to ensure that it performs as needed. As a result, there are two parts to the prototype system testing and assessment in this study. These include user acceptance testing and system performance testing.

### 6.4.1 System Performance Testing

It is the process of determining whether it achieves the required level of accuracy or not and the system's accuracy is evaluated using the test cases [58]. The test cases include samples of disease instances taken from the Werabe Referral Hospital dataset. This testing method is performed to asses or evaluates the performance of the prototype system using the parameter precision, recall, F-measure, and accuracy. The researcher choose 21 test cases in all are prepared for system performance evaluation. There are 10 test cases for Severe, 6 test cases for Mild and 5 for Moderate. Domain experts receive the test cases, which are instances of unidentified diseases, and label them as severe, mild and moderate. The knowledge-based system is given this set of test cases, and its outputs are compared against the opinion of the domain expert. Table 6.2 below shows the confusion matrix of the three classes.

Table 6.2: Confusion Matrix for System Performance Testing

| **Domain expert Judgment** | **Knowledge base Judgment** | | | | |
|---|---|---|---|---|---|
| | Class | Severe | Mild | Moderate | Total |
| | Severe | 9 | 1 | 0 | 10 |
| | Mild | 0 | 6 | 0 | 6 |
| | Moderate | 1 | 0 | 4 | 5 |
| | Total | 10 | 7 | 4 | 21 |

The above confusion matrix shows test case evaluation by the developed KBS and domain expert judgment. The value inside the severe row shows that out of 10 instances the system identified **9** as severe correctly and 1incorrectly classified as Moderate. The value inside the Mild row shows that out of **6** instances the system identified **6** as Mild correctly. The value inside the Moderate row shows that out of **5** instances the system identified **4** as moderate correctly. In general, from 21 diagnosed patient test cases **19** cases diagnosed patients' test cases are classified correctly and 2 diagnosed patient test cases are classified incorrectly. The test case result that provided by system evaluators showed that the prototype system is about **90.5** % diagnosis accuracy for Hospital Acquired Pneumonia Disease.

As presented in Table 6.3, the system performance is evaluated in terms of True Positive Rate (Recall), Precision, F-measure, and False Positive Rate. Table 6. 3: The Detailed Accuracy of System Performance Testing.

Table 6.3: Accuracy of System Performance Testing

| | **TPR** | **Precision** | **FPR** | **F-measure** | **Class** |
|---|---|---|---|---|---|
| | 0.89 | 0.88 | 0.08 | 0.89 | Severe |
| | 1 | 0.86 | 0.06 | 0.95 | Mild |
| | 0.84 | 1 | 0 | 0.89 | Moderate |
| **AVG Weight** | 0.91 | 0.913 | 0.05 | 0.91 | |

As shown in the table above the system scores a weight average of 91% TPR, 91.3% precision, 91% F-measure and 5 % FPR to diagnose and recommend the treatment of Hospital Acquired Pneumonia. Based on system performance testing the system achieves 91.3% accuracy.

### 6.4.2 User Acceptance Testing

User acceptability testing aims to determine whether the system is adequate and accurate for usage, as well as whether it can support daily activities and user scenarios. All of the effort might be useless if customers reject the KBS. Thus, it is an essential task for realizing a

knowledge-based system. Four domain specialists who had worked in the Werabe Referral Hospital for a considerable amount of time and were willing to assess the system were chosen for user acceptance testing. The following evaluation standards were applied by the domain experts to evaluate the prototype system's performance.

Various researchers have employed various evaluation criteria for user acceptance evaluation of the system, the criteria for this studies are taken from [35] These criteria are

- Simplicity of the system to use
- Attractiveness of the system
- Efficiency in time
- The accuracy of the system in reaching a decision to identify the severity level of the diseases.
- The ability of the system in making the right treatment recommendations
- Importance / contribution of the system in the domain area.

We determined values for every assessment criterion on the checklist so that domain experts could assess how well the prototype system performed. Thus, the researcher assigned a value to each word within the scale in order to examine the system performance with user evaluations. Domain experts are put their values based on the given Likert scale as Excellent = 5, Very good = 4, Good = 3, Fair = 2 and Poor = 1 for each criterion of evaluation. The user acceptance of the system is measured manually as follow.

$$\text{AvgS} = SV1 * \frac{NR1}{\overline{TNR}} + SV2 * \frac{NR2}{\overline{TNR}} \ldots \sum_{i=1} SVi * \frac{NR}{\overline{TNR}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6.1)$$

Where, AvgS average score, SV scale value, TNR total number of respondents and NR is the number of respondents. To get the result of user acceptance average performance is calculated out 100%.

$$\text{AvgP} = \frac{AvgS}{NS} * 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6.2)$$

Where, NS is the Number of Scales and AvgP is Average Performance and AvgS is average score. The table 6.4 below summarizes the results obtained from the respondent.

Table 6.4: User Acceptance Evaluation Criteria and their Result

| No | Criteria of evaluation | Poor (1) | Fair (2) | Good (3) | Very good (4) | Excellent (5) | Average score | Average performance % |
|---|---|---|---|---|---|---|---|---|
| 1 | Simplicity of the system to use | 0 | 0 | 0 | 2 | 3 | 4.6 | 92 |
| 2 | Attractiveness of the system | 0 | 0 | 1 | 3 | 1 | 4 | 80 |
| 3 | Efficiency in time | 0 | 0 | 1 | 0 | 4 | 4.6 | 92 |
| 4 | Ability of the system in making the right treatment recommendations | 0 | 0 | 0 | 2 | 3 | 4.6 | 92 |
| 5 | Accuracy of the system in reaching a decision to identify the severity level of the diseases | 0 | 0 | 0 | 0 | 5 | 5 | 100 |
| 6 | Contribution of the system in the domain area. | 0 | 0 | 1 | 0 | 4 | 4.6 | 92 |
| | | | | | | | 4.9 | 91.3 |

As can be seen in table 6.3 for simplicity of the system to use, 2 evaluators rated the system as very good and 3 evaluators rated as excellent. It can be interpreted as from the total evaluators, 40% of the respondents evaluated as good and 60% of the respondents evaluated as Excellent. For the second criteria or the attractiveness of the system 1 evaluator rated as good, 3 evaluators rated as very good and 1 as excellent. It can be interpreted as 20% of the respondent evaluated as good, 60% of the respondent ticked as very good and the remaining 20%of them evaluated as excellent. For the third criteria that is efficiency of the system in time1 evaluators rated as

good and 4 evaluators rated as excellent. It can be interpreted as from the total evaluators, 20%of the respondents evaluate the system as good and 80% of the respondent evaluated as excellent.

For the fourth criteria that is ability of the system in making the right treatment recommendation 2 evaluators rated as very good and 4 evaluators as excellent. And it can be interpreted as from a total of evaluators 40% of the evaluator rated as very good and 60 % of them evaluated as excellent. For the fifth criteria 5 evaluators rated as excellent .it is interpreted the Accuracy of the system in reaching a decision to identify the severity level of the diseases, all evaluators rated the system as Excellent means from the total evaluators 100% rated as Excellent. The average of the evaluator's result shows that the system is 100% efficient in terms of time. For the last criteria that is the contribution of the system in the domain area 1 evaluator rated as good 4 evaluators as excellent this means that 20 % of the evaluators evaluated as good and the remaining 80%of the respondent evaluated as excellent.

In conclusion, the domain experts' evaluation results show that the system performs on average 4.9 out of 5, or 91.3%, above very good.

## 6.5 Discussion of Result

This study was developed using manual and automatic knowledge acquisition approaches and the extracted knowledge has collaborated. So, it makes a difference from a previous study which is listed in section 2.8. Based on the result obtained, in this section, the researcher would discuss the result acquired concerning the study objective and research questions itemized in chapter one. At the beginning of this study, three research questions were encompassed. Hence, this section tries to answer those questions.

To select the determinant attributes, gain ratio attribute evaluator/selection techniques and domain expert’s interview are used and the study finds out that Chills, Increased Vocal Fremitus (InVoFr), Dyspnea, chest pain, Dullness to percussion (DtoP), sex, Vomiting, shortness of breath (SoFBreath), white blood cell (AbWBC), malaise, fever, tachypnea, Blood Pressure (DecBP), LoAppetite and Temperature are the determinant attributes that are used to diagnose the selected Hospital Acquired Pneumonia.

The researcher selected appropriate classification algorithm that was used to build the classifier model for HAP diagnosis. There are four experiments with two scenarios for each algorithm are made, the first scenario is with all attributes and with selected attributes using four classification algorithms these are J48, JRip, PART and Random Forest under 10fold cross- validation, and percentage split test option. Based on section 5.5, from four experiments PART algorithm with conducted on selected attribute with percentage split test options were chosen with an accuracy of **99.3%** as the best classifier model for the development of KBS.

The knowledge for developing the Knowledge Base System (KBS) is extracted from the collected dataset and domain experts. The best rule is extracted by applying the PART rule induction classifier algorithm compared to other classification algorithms based on performance evaluation metrics. To acquire knowledge from expert the researcher used semi-structured interview techniques. After the knowledge is acquired, it has been modeled in appropriate techniques. Generally, the knowledge acquired from collected data and from experts is listed in Appendix III.

# CHAPTER SEVEN

# CONCLUSION AND RECOMMENDATION

The researcher attempted to provide a summary of the study, identify the study's benefits to the organization and healthcare professionals, and make pertinent recommendations for the organizations and future work in this chapter.

## 7.1 Conclusion

Hospital-acquired (nonsocial) pneumonia (HAP) is a pneumonia that occurs 48 hours or more after admission, Alveoli, which are tiny air sacs in the lungs; fill with air when a healthy individual breathes. When someone has pneumonia, their alveoli are packed with pus and fluid, which makes breathing challenging and lowers oxygen intake.

This infection can be contracted outside of a medical setting as well. It can also be spread through inhaled or aspirated bacteria. Additionally, it is a major global health issue that contributes significantly to be the first in morbidity and mortality. In the medical field, data mining techniques are typically integrated with rule-based or case-based reasoning systems [59].

In this study, the researcher has developed knowledge-based system (KBS) for diagnosis and treatment recommendation of Hospital Acquired Pneumonia (HAP) by collaborating data mining result with expert knowledge. The dataset is gathered from the pediatric ward of Werabe Hospital, and preprocessed using data mining tools. To achieve the healthcare service goal, the developed KBS plays an important role in improving the healthcare facilities, increasing the efficiencies of a healthcare institution, to reduce the economic challenges in the healthcare management process, and reduce the mortality rate.

The Design Science Research approach is the selected methodology of this study that grows from relevance and rigor and hybrid data mining process model is used. Therefore, the researcher has made an effort to follow relevance of identifying problems, opportunities that exist in healthcare and applicable knowledge that allows us to develop an artifact from rigor. For the data mining process, the researcher has collected 3244 instances from Werabe referral Hospital. The instances have 21 attributes and 3 class labels. The class labels are Severe, Mild

and moderate. After the data has been collected, it is preprocessed using preprocessing techniques and prepared for the format suitable for data mining. Under the data preprocessing, three tasks were performed these are data cleaning, handling missing value, data transformation, and attribute selection. From the total attributes, 16 attributes are selected as a determinant for the diagnosis and treatment recommendation of Hospital Acquired Pneumonia. The attribute selection was done by using gain ratio and domain expert guidance.

Four classification algorithms are used for building the classifier model; these are J48, JRip, PART and Random Forest algorithm. And then four experiments were performed with selected algorithms and each experiment has two scenarios. Those are conducted with whole and selected attributes under 10-fold cross-validation and percentage split test options. After developing the model, the researcher has made a performance comparison of each model, PART classifier algorithm conducted on selected attributes with percentage split test option with an accuracy of **99.3%** was achieved. After the knowledge is extracted from the data mining the next is acquiring knowledge from the domain expert. Semi –structured interview technique is chosen for acquiring knowledge from expert. After the knowledge was acquired, it is modeled by using the decision tree modeling techniques and represented in the production rule. The two extracted knowledge was combined and checked for rule redundancy to develop the knowledge-based system.

To develop the KBS the researcher used SWI prolog and Net Beans for making user interface. The Developed Knowledge-based system has Knowledge Base, Inference Engine, Explanation Facility, and User Interface components. To evaluate performance of the developed system, the researcher has used system performance testing by preparing test cases and user acceptance evaluation. 21 test cases to evaluate the performance of the system by comparing it with an expert's judgment are used and it achieves **90.5%** accuracy. Then user acceptance testing is performed based on seven criteria of evaluation. Selected domain experts are trained and used the system to evaluate the developed knowledge base system requirements. Based on the user's evaluation the system scores **91.3%** of accuracy. The result show that the developed system achieves good performance and meets the objectives of the study and it could give proper treatment.

## 7.2 CONTRIBUTION OF THE STUDY

In order to acquire knowledge, prepare data for preprocessing, conduct experimental analysis, and create a knowledge-based system, this research has carefully examined the theory, methodology, and basis that now exist on the diagnosis and treatment of Hospital Acquired Pneumonia (HAP). In order to meet the goals, this study has produced a number of results that add to the body of knowledge and appropriate environment. This study's main contribution is an unusual knowledge-based system artifact design, or model, that solves a real-world issue in healthcare settings.

The outcome of this research is a knowledge-based system that could be used to identify HAP illnesses and recommend an appropriate course of treatment. Collaboration of data mining result with knowledge of an expert is another contribution of this research. The constructed artifact helps medical professionals detect illnesses more quickly, lowers down on unnecessary diagnostic procedures, and serves as a guide for aspiring doctors who must diagnose and provide therapies and assist them in decision making process.

## 7.3 RECOMMENDATION

Based on the findings of this study, the following recommendations are suggested for health organization and future work.

### 7.3.1 Recommendation for the healthcare organization

Hospital Acquired Pneumonia was successfully diagnosed by the developed technique, and treatment recommendations were made. Because of the developed system's larger benefits to the organization and its employees, the researcher strongly advises healthcare organizations to implement it for improving the experience of health professionals.

### 7.3.2 Recommendation for future work

- The study is developed by collaborating data mining result with expert’s knowledge. This developed knowledge is based on rule-based reasoning system. To increase the current work performance further research can be done by collaborating data mining results with a hybrid reasoning system (rule-based and case-based reasoning system) techniques.

- Other than speaking in Amharic with patients, healthcare professionals in the health center perform a number of tasks in English. Moreover, designing and developing a self-learning Amharic based Knowledge based system (KBS) that can provide advice through Amharic or local language user interface for physician and patients in order to facilitate the diagnosis and treatment of Hospital Acquired Pneumonia (HAP) should be considered as future task.
- Building a knowledge-based system for the diagnosis and treatment recommendation of Hospital Acquired Pneumonia (HAP) was the main goal of this research, along with determining the determinant elements that are employed for the diagnosis of the diseases. As a result, more research on pneumonia-like illnesses in the future, such as ventilator-associated pneumonia, is required. Additionally, comprehensive research is required for the diagnosis and timely provision of recommendations.
- As future work using image processing to create models and gather image data using deep learning algorithm will improve the accuracy of diagnosing Hospital Acquired Pneumonia (HAP).

## REFERENCES


[1] G. Mackenzie, 'The definition and classification of pneumonia', *pneumonia*, vol. 8, no. 1, pp. 14, s41479-016-0012-z, Dec. 2016.

[2] V. Chouhan *et al.*, 'A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images', *Applied Sciences*, vol. 10, no. 2, p. 559, Jan. 2020.

[3] Z. Fatmi and F. White, 'A comparison of "cough and cold" and pneumonia: risk factors for pneumonia in children under 5 years revisited', *International Journal of Infectious Diseases*, vol. 6, no. 4, pp. 294–301, Dec. 2002.

[4] Faculty of Computing, Bahir Dar University, Ethiopia, A. Nega, and A. Kumlachew, 'Data Mining Based Hybrid Intelligent System for Medical Application', *IJIEEB*, vol. 9, no. 4, pp. 38–46, Jul. 2017.

[5] E. Tsegaye, 'Datamining result based hybrid knowledge-based system for pediatric community acquired pneumonia diagnosis and treatment recommendation '2021.

[6] A. Mishra, B. B. Gupta, D. Peraković, and F. J. G. Peñalvo, 'A Survey on Data mining classification approaches'2021.

[7] N. T. Sharew, H. T. Bizuneh, H. K. Assefa, and T. D. Habtewold, 'Investigating admitted patients' satisfaction with nursing care at Debre Berhan Referral Hospital in Ethiopia: a cross-sectional study', *BMJ Open*, vol. 8, no. 5, p. e021107, May 2018, doi: 10.1136/bmjopen-2017.

[8] M. J. Choi *et al.*, 'Disease burden of hospitalized community-acquired pneumonia in South Korea: Analysis based on age and underlying medical conditions', *Medicine*, vol. 96, no. 44, p. e8429, Nov. 2017.

[9] Y. Li, Z. Zhang, C. Dai, Q. Dong, and S. Badrigilan, 'Accuracy of deep learning for automated detection of pneumonia using chest X-Ray images: A systematic review and meta-analysis', *Computers in Biology and Medicine*, vol. 123, p. 103898, Aug. 2020.

[10] M. S. Lee *et al.*, 'Guideline for Antibiotic Use in Adults with Community-acquired Pneumonia'2018.

[11] C. Diriba, M. Meshesha, and D. Tesfaye, 'Developing a Knowledge-Based System for Diagnosis and Treatment of Malaria', *J. Info. Know. Mgmt.*, vol. 15, no. 04, p. 1650036, Dec. 2016.

[12] L. T. Daminova, N. Z. Asadov, and D. K. Muminov, 'Out-Of-Social Pneumonia On The Background Of Chronic Kidney Disease', *Clinical Medicine*, vol. 07, no. 03, 2020.

[13] A. R. Modi and C. S. Kovacs, ‘Hospital-acquired and ventilator-associated pneumonia: Diagnosis, management, and prevention’, *CCJM*, vol. 87, no. 10, pp. 633–639, Oct. 2020.

[14] D. A. McAllister *et al.*, ‘Global, regional, and national estimates of pneumonia morbidity and mortality in children younger than 5 years between 2000 and 2015: a systematic analysis’, *The Lancet Global Health*, vol. 7, no. 1, pp. e47–e57, Jan. 2019.

[15] A. Agweyu *et al.*, 'Appropriateness of clinical severity classification of new WHO childhood pneumonia guidance: a multi-hospital, retrospective, cohort study', *The Lancet Global Health*, vol. 6, no. 1, pp. e74–e83, Jan. 2018.

[16] D. R. House, S. Rijal, S. Adhikari, M. L. Cooper, and C. M. Hohl, 'Prospective evaluation of World Health Organization guidelines for diagnosis of pneumonia in children presenting to an emergency department in a resource-limited setting', *Paediatrics and International Child Health*, vol. 40, no. 4, pp. 227–230, Oct. 2020.

[17] Elina Naydenova , Athanasios Tsanas , Stephen Howie 'Neonatal Intensive Care Unit (NICU) Training'2017.

[18] U. Shafique and H. Qaiser, 'A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)', vol. 12, no. 1, 2014.

[19] U. Shafique and H. Qaiser, 'A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)', vol. 12, no. 1, 2014.

[20] E. F. Olana, 'Predicting Maternal Mortality Rate Using Data Mining Techniques: The Case of Jimma University Specialized Hospital Maternity Wards', vol. 3, no. 2, 2022.

[21] M. Shah and S. Nair, 'A Survey of Data Mining Clustering Algorithms', *IJCA*, vol. 128, no. 1, pp. 1–5, Oct. 2015.

[22] S. T. March and G. F. Smith, 'Design and natural science research on information technology', *Decision Support Systems*, vol. 15, no. 4, pp. 251–266, Dec. 1995.

[23] Assoc. Professor, CVR College of Engineering/CSE Department, Hyderabad, India, S. Nimmala, D. Sujan Kumar, and Assoc. Professor, CVR College of Engineering/CSE Department, Hyderabad, India, 'High Blood Pressure Prediction based on AAA++ using J48 Algorithm', *CVRJST*, vol. 14, no. 01, pp. 53–57, Jun. 2018.

[24] J. Hussain and S. Lalmuanawma, 'Feature Analysis, Evaluation and Comparisons of Classification Algorithms Based on Noisy Intrusion Dataset', *Procedia Computer Science*, vol. 92, pp. 188–198, 2016.

[25] D. L. A. AL-Nabi and S. S. Ahmed, 'Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)', 2013.

[26] Z. Markos, 'Predicting Under Nutrition Status of Under-Five Children Using Data Mining Techniques: The Case of 2011 Ethiopian Demographic and Health Survey', *J Health Med Informat*, vol. 5, no. 2, 2014.

[27] V. Matzavela and E. Alepis, ‘Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments’, *Computers and Education: Artificial Intelligence*, vol. 2, p. 100035, 2021.

[28] P. Bhakare and D. S. Suraj, ‘Data Mining in Healthcare: Current Applications and Issues’, vol. 11, no. 11, 2022.

[29] H. Kaur and S. K. Wasan, ‘Empirical Study on Applications of Data Mining Techniques in Healthcare’, *J. of Computer Science*, vol. 2, no. 2, pp. 194–200, Feb. 2006.

[30] M. A. Bramer, *Principles of data mining*. in Undergraduate topics in computer science. London: Springer, 2007.

[31] A. Arooj, M. Riaz, and M. N. Akram, ‘Evaluation of predictive data mining algorithms in soil data classification for optimized crop recommendation’, in *2018 International Conference on Advancements in Computational Sciences (ICACS)*, Lahore, Pakistan: IEEE, Feb. 2018.

[32] D. Aweke Wako, ‘Development of Knowledge Based System for Wheat Disease Diagnosis: A Rule Based Approach’, *J Comput Eng Inf Technol*, vol. 06, no. 05, 2017.

[33] K. P. Tripathi, ‘A Review on Knowledge-based Expert System: Concept and Architecture’, *Artificial Intelligence Techniques*, 2011.

[34] R. G. Smith, ‘Knowledge-Based Systems: Concepts, Techniques, Examples’2023.

[35] College of Engineering and Technology, Jigjiga University, Ethiopia and S. Mohammed, ‘A Self-learning Knowledge based System for Diagnosis and Treatment of Chronic Kidney Disease’, *IJEME*, vol. 9, no. 2, pp. 44–58, Mar. 2019.

[36] T. Aboneh, ‘knowledge based system for pre-medical triage treatment at adama university asella hospital’2013.

[37] S. Yitagesu, Z. Feng, M. Meshesha, G. Mekuria, and M. Q. Yasin, ‘Developing Knowledge-Based Systems Using Data Mining Techniques for Advising Secondary School Students in Field of Interest Selection’, in *Database Systems for Advanced Applications*, vol. 10829, C. Liu, L. Zou, and J. Li, Eds., in Lecture Notes in Computer Science, vol. 10829. , Cham: Springer International Publishing, 2018.

[38] R. Plant and R. Gamble, ‘Methodologies for the development of knowledge-based systems, 1982–2002’, *The Knowledge Engineering Review*, vol. 18, no. 1, pp. 47–81, Jan. 2003.

[39] E. L. Rissland and D. B. Skalak, ‘Combining Case-Based and Rule-Based Reasoning: A Heuristic Approach’2009.

[40] Restya Winda Astari, C. H. Ayuningtyas, and G. A. Putri Saptawati, ‘Knowledge based system for supporting genomic based personalized medicine’, in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, Bandung, Indonesia: IEEE, Jul. 2011.

[41] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, ‘A Design Science Research Methodology for Information Systems Research’, *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, Dec. 2007.

[42] E. Naydenova, A. Tsanas, S. Howie, C. Casals-Pascual, and M. De Vos, ‘The power of data mining in diagnosis of childhood pneumonia’, *J. R. Soc. Interface.*, vol. 13, no. 120, p. 20160266, Jul. 2016.

[43] V. Krishnaiah, D. G. Narsimha, and D. N. S. Chandra, ‘Survey of Classification Techniques in Data Mining’, *International Journal of Computer Sciences and Engineering*, vol. 2, 2014.

[44] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. in Intelligent Systems Reference Library, vol. 72. Cham: Springer International Publishing, 2015.

[45] J. Hussain and S. Lalmuanawma, ‘Feature Analysis, Evaluation and Comparisons of Classification Algorithms Based on Noisy Intrusion Dataset’, *Procedia Computer Science*, vol. 92, pp. 188–198, 2016.

[46] M. Assefa and M. Meshesha, ‘A Combined Reasoning System for Knowledge Based Network Intrusion Detection’, vol. 7, no. 1, 2019.

[47] S. Ozarslan and P. E. Eren, ‘MobileCDP: A mobile framework for the consumer decision process’, *Inf Syst Front*, vol. 20, no. 4, pp. 803–824, Aug. 2018.

[48] N. Padhy, ‘The Survey of Data Mining Applications and Feature Scope’, *IJCSEIT*, vol. 2, no. 3, pp. 43–58, Jun. 2012.

[49] G. Quin, ‘Chest pain evaluation units’, *Emergency Medicine Journal*, vol. 17, no. 4, pp. 237–240, Jul. 2000.

[50] ‘Diagnosis of hospital-acquired pneumonia and methods of testing for pathogens’, *Respirology*, vol. 14, no. s2, Nov. 2009.

[51] R. Wirth and J. Hipp, ‘CRISP-DM: Towards a Standard Process Model for Data Mining’2000.

[52] S. Raschka, ‘Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning’. arXiv, Nov. 10, 2020. Accessed: Apr. 04, 2024. [Online].

[53] J. Jackson, ‘Data Mining; A Conceptual Overview’, *CAIS*, vol. 8, 2002.

[54] F. Alonso, J. P. Caraça-Valente, A. L. González, and C. Montes, ‘Combining expert knowledge and data mining in a medical diagnosis domain’, *Expert Systems with Applications*, vol. 23, no. 4, pp. 367–375, Nov. 2002.

[55] E. Lughofer and M. Sayed-Mouchaweh, ‘Prologue: Predictive Maintenance in Dynamic Systems’, in *Predictive Maintenance in Dynamic Systems*, E. Lughofer and M. Sayed-Mouchaweh, Eds., Cham: Springer International Publishing, 2019.

[56] B. A. Beemer and D. G. Gregg, ‘Dynamic interaction in knowledge based systems: An exploratory investigation and empirical evaluation’, *Decision Support Systems*, vol. 49, no. 4, pp. 386–395, Nov. 2010.

[57] J. Ismail, ‘The design of an e-learning system Beyond the hype’, *Internet and Higher Education*, 2002.

[58] N. Kerdprasop and K. Kerdprasop, ‘Autonomous Integration of Induced Knowledge into the Expert System Inference Engine’, *Hong Kong*, 2011.

[59] S. Vanaja and K. Rameshkumar, ‘Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey’, *Journal of Computer Science*, vol. 11, no. 1, pp. 30–52, Jan. 2015.

## APPENDIX I:

User acceptance Testing evaluation criteria

Dear Evaluator,

This assessment questionnaire was created with the intention of determining the extent to which end users in the vicinity of the health facility find the diagnosis and treatment of hospital acquired pneumonia (HAP) KBS acceptable and useful.

As a result, you are cordially asked to assess the system by marking the (√) symbol in the designated spot for the associated attribute values for every assessment criterion. I want to thank you for your assistance in giving the data. The values are rated as: **Excellent**=5, **Very good** =4, **Good**=3, **Fair**= 2 and **Poor** =1

| **No** | **Criteria of evaluation** | **Poor (1)** | **Fair (2)** | **Good (3)** | **Very good (4)** | **Excellent (5)** | **Average score** | **Average performance %** |
|---|---|---|---|---|---|---|---|---|
| 1 | Simplicity of the system to use | 0 | 0 | 0 | 2 | 3 | 4.6 | 92 |
| 2 | Attractiveness of the system | 0 | 0 | 1 | 3 | 1 | 4 | 80 |
| 3 | Efficiency in time | 0 | 0 | 1 | 0 | 4 | 4.6 | 92 |
| 4 | Ability of the system in making the right treatment recommendations | 0 | 0 | 0 | 2 | 3 | 4.6 | 92 |
| 5 | Accuracy of the system in reaching a decision to identify the severity level of the diseases | 0 | 0 | 0 | 0 | 5 | 5 | 100 |
| 6 | Contribution of the system in the domain area. | 0 | 0 | 1 | 0 | 4 | 4.6 | 92 |
| | | | | | | | **4.9** | **91.3** |

## APPENDIX II:

## Dear interviewees;

I want to start by expressing my gratitude for accepting to meet with you for the interview. From the University of Wolkite, I'm Wondimu Kibatu. As a postgraduate student in my second year, I am working with expert knowledge to combine data mining findings with knowledge from other fields to produce a knowledge-based system for the diagnosis and treatment recommendations of Hospital Acquired Pneumonia. Given that the topic of this thesis is "**Developing Classification Model with Knowledge Base System for Diagnosis and Treatment of Hospital Acquired Pneumonia**," you are likely familiar with this disease.

1. What is pneumonia?
2. Which pneumonia diseases type is frequently occurred in this hospital?
3. What are the symptoms of each Hospital Acquired Pneumonia?
4. How to diagnose Hospital Acquired Pneumonia?
5. What type of treatments is given for this disease?

## APPENDIX III:

**Rule 1:** IF fever =No AND Malaise= Yes AND Temperature=High AND InVoFr=Yes AND Chest Pain =Stabbing AND Vomiting =yes: THEN it is Severe (12.0/1.0)

**Rule 2:** IF InVoFr =Yes AND Fever =Yes AND Malaise =No AND Sex =Male AND DecBP= Diastolic AND Temperature =Normal AND AbWBC= Leukopenia: THEN it is Moderate (3.0).

**Rule 3:** IF Malaise= Yes AND Tachypnea = Yes AND Dyspnea =No AND DtoP = Hyper rmonant AND AbWBC =Leukopenia: THEN it is Mild (7.0).

**Rules extracted from expert for the diagnosis of Hospital Acquired Pneumonia**

**Rule 1:** IF patient has Nausea, THEN the diseases =Severe HAP.

**Rule 2:** IF patient has No Nausea AND Muscle Fatigue Yes THEN the diseases =Mild HAP.

**Rule 3:** IF patient has No Nausea with muscle fatigue No and breathing difficulty is Yes THEN the diseases =Moderate HAP.

**Rule 4:** IF patient has Fever, THEN the diseases =Moderate HAP.

**Rule 5:** IF patient has No fever with Feeling very tiered is Yes, THEN the disease =Severe HAP.

**Rule 6:** IF patient has vomiting THEN the diseases =Severe HAP.

**Rule 7:** IF patient has No vomiting AND chills Yes THEN the diseases =Mild HAP.

**Rules extracted from expert for the treatment of Hospital Acquired Pneumonia**

**Rule 1:** IF the diagnosis results of HAP =Severe, THEN the treatment= Piperacillin /tazobactam 4.5 g IV 8 hourly plus Gentamicin5 mg/kg/day IV, Beta-Lactam antibiotic, Piperacillin.

**Rule 2:** IF the diagnosis results of HAP =Mild, THEN the treatment= Cephazolin 1 g IV 8 hourly plus Gentamicin* 5 mg/kg/day IV.

**Rule 3:** IF the diagnosis results of HAP =Moderate, THEN the treatment=amoxycillin + clavulanic acid 875/125 mg (1 tablet) orally 12 hourly.