DEBRE BIRHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF INFORMATION SYSTEMS

Integrating Data Mining Results with Knowledge Based System for Diagnosis and Treatment of Cattle Diseases: The Case of Debre Birhan Basso Animal Health Center

By

TADESSE BEYENE GASHAWTENA

Jun. 2018

DEBRE BIRHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF INFORMATION SYSTEMS

Integrating Data Mining Results with Knowledge Based System for Diagnosis and Treatment of Cattle Diseases: The Case of Debre Birhan Basso Animal Health Center

A Thesis Submitted to the College of Graduate Studies of Debre Berhan University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Systems

By

TADESSE BEYENE GASHAWTENA

Jun. 2018

Debre Birhan University

Debre Birhan, Ethiopia

DEBRE BIRHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF INFORMATION SYSTEMS

Integrating Data Mining Results with Knowledge Based System for Diagnosis and Treatment of Cattle Diseases: The Case of Debre Birhan Basso Animal Health Center

By

TADESSE BEYENE GASHAWTENA

Name and Signature of Members of the Examining Board

Name	Title	Signature	Date
Dr. Kindie Biredagn	Advisor	_______	__________
Dr. G. Mohamed Gouse	Chair Person	_______	__________
Dr. Tibebe Beshah	External Examiner	_______	__________
Dr. Nilamadhab Mishra	Internal Examiner	_______	__________

DECLARATION

I am Tadesse Beyene, a student in the College of Computing, University of Debre Birhan, aware of my responsibility, the penal law, declare and certify with my signature that my thesis entitled Integrating Data Mining Results with Knowledge Based System for Diagnosis and Treatment of Cattle Diseases: The Case of Debre Birhan Basso Animal Health Center is entirely the result of my own work. I have faithfully and accurately cited, all my sources, including books, journals, handouts and unpublished manuscripts, as well as any other media, such as the Internet.

I understand that

- ✓ Literal citing without using quotation marks and marking the references
- ✓ Citing the contents of a work without marking the references
- ✓ Using the thoughts of somebody else whose work was published, as of our own thoughts counts as plagiarism.

I declare that I understood the concept of plagiarism and I acknowledge that my thesis will be rejected in case of plagiarism.

Declared By:

Name: Tadesse Beyene

Signature: ______________

Date: _______________

Confirmed By:

Name: Dr. Kindie Biredagn (Advisor)

Signature: ______________

Date: ______________

DEDICATION

I would like to dedicate my thesis to my father, who passed away without seeing my achievement, and to my dear brother, Bayu Beyene who devoted his life for my achievement. Thank you Emamey! God bless your life in the best of her ways.

ACKNOWLEDGEMENT

First and foremost, praises and thanks to the **God**, the Almighty, for His showers of blessings throughout my research work to complete the research successfully.

I would like to thank my advisor, **Dr. Kindie Biredagn**, for his guidance, direction, patience, and continuous encouragement throughout the time of my thesis research. He was always there to answer my questions and set time to meet with me whenever needed. I will be forever grateful.

I would like to express my respectful thanks to my family, especially my Beloved Fiancé **Etag** and her brother **Dejene Shewakena** for their words of encouragement and support are what kept me going all these years. I sincerely thank my whole family, without their support and encouragement I would have not been here today. They have always been with me supporting, helping and appreciating in my journey to be a better man.

I would like to thank Basso Animal Health Center Workers, **Dr. Hailemariam Amare, Mulatu Tegegn, Tsegu Teklyie, Fantahun Gebiru, Yonas Tadele** and administration of Debre Berhan animal health officer **Mr. Eshetu Demissie**. I always wish for good as half you are.

I would like to express my gratitude for Debre Birhan University College of computing members, especially **Yeshalem Gezahegn, Tigist Adem, Kindie Alebachew, Mikre Getu (Mak), Baye Yemataw, Hagos Yirgaw** and **Almaz Assefa**. And also I would like to DBU main library workers specially, **Tekletsadik Aklilu** and **Belete Shimels**.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Table of Contents

List of Tables

List of Figures

Table of Lists

Acronym

AI	Artificial Intelligence
ARFF	Attribute Relation File Format
BAHC	Basso Animal Health Center
C/I	Contraindication
CART	Classification & Regression Trees
CBR	Case Based Reasoning
CommonKADS	Common Knowledge Acquisition and Documentation Structuring
CRISP-DM	CRoss Industry Standard Process for Data Mining
CSV	Comma Separated Value
D/F	Dosage Forms
D/I	Drug Interaction
DBAHC	Debre Birhan Animal Health Center
FMD	Foot Mouse Disease
GUI	Graphical User Interface
ID3	Iterative Dichotomiser 3
IIS	Internet Information Server
IM	Intramuscularly
IU	International Unit
IV	Intravenously
JDE	Java Integrated Development Environment
JDK	Java Development Kit
JPL	Java to Prolog Library
KBS	Knowledge Based System
KDD	Knowledge Discovery Database
KDP	knowledge discovery processes
KR	Knowledge Representation
LSD	Lumpy skin Disease
MSF	Medicine Sans Frontieres
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
SMOTE	Synthetic Minority Oversampling Technique
S/E	Side Effect
SEMMA	Sample, Explore, Modify, Model Asses
SWI-PROLOG	PROGraming in Logic
VMD	Veterinarian Medicine Doctors
WEKA	Waikato Environment for Knowledge Analysis
WP	Withdrawal Period

Abstract

The general objective of this study is to construct predicting model and integrate automatically with the knowledge base system, in order to increase the efficiency in the diagnosis and treatment of common cattle disease. Different researches, articles, journal papers and guidelines that have been done on cattle disease diagnosis and treatment, data mining, KBS, knowledge acquisition for KBS and integration of data mining results with KBS were reviewed. Besides of this, related works were reviewed to identify the gap and formulate research questions of the study. Mixed (quantitative and qualitative) research design was used, integrated (manual and automated) knowledge acquisition techniques were used to acquire knowledge, rule based knowledge representation approach was used to represent knowledge in the knowledge base.

Away from each researchers and as to the review of the researcher, no study was done at the area throughout the Amhara Region animal health center using data mining techniques. For this study, the researcher see that the limitation of the existing system is not considering analysis of data in detail.

In this study, rule based diagnosis and treatment of cattle disease knowledge based system is proposed. The system is aimed at utilizing hidden knowledge extracted by employing an induction algorithm of data mining, specifically JRip from the sampled BAHC dataset. The Integrator application, then links the model created by JRip classifier to knowledge based system so as to add knowledge automatically. In doing so, the Integrator understands the syntax of JRip classifier and Prolog and converts from rule representation in JRip to Prolog understandable format.

To do this, java programming was used to integrate WEKA result with the Knowledge Based System automatically. And also ‘swiweka’ is used as an interface that allows the use of WEKA API for classification; weka.jar, Weka _src.jar are used to construct a model when called from interface through swiweka package, JPL library to connect the Java layer with the Prolog layer.

Finally, as a result, the proposed system can perform in the absence of domain experts with 98.7% accuracy of recall evaluation results which indicates that the KBS has the necessary knowledge for diagnosis and treatment of cattle disease which in turn implies that the study was effective in acquiring knowledge. Besides of this, the proposed system achieves 85.8% of the users' acceptance which in turn implies that the proposed system could be operational if it could be implemented.

Keywords: - Cattle disease, Data mining, rule based, integrated with the Knowledge Base System, JRip classifier, KBS, CommonKADS, SWIWEKA.

CHAPTER ONE

INTRODUCTION

1.1. Background

Ethiopia is one among the nations that possesses the largest livestock population in the African continent with an estimated fifty six Million of cattle, fifty eight Million of sheep and goats and ten Million of equines, one Million of camels and fifty seven Million of chicken [1]. The average lifespan for cattle is 18 to 22 years, although they can live in excess of 25 years. Because many rescued animals have come from abusive conditions, however, these cattle may have more health problems and a shorter life span than other cattle [2].

Livestock sector would be a significant contributor to Ethiopia's economy. It is noteworthy livestock products and by products like meat, milk and butter that contribute to the improvement of the nutritional status of the people. Livestock also plays an important role in providing export commodities, such as live animals and skins, which benefits the country with foreign exchanges. Livestock also gives a certain degree of security in times of crop failure, as they are a near-cash capital stock. Furthermore, livestock provides farmyard manure that is commonly applied to improve soil fertility and also used as a source of energy.

In addition to death, they cause loss of production and frequently a loss of body condition [3]. Unhealthy animals require more food and take longer time for growth than healthy ones. The cattle disease costs Millions of Birr losses every year. Generally, animals are born free of disease or parasites. But they usually acquire these disorders either through contact with diseased animals or due to improper sanitation, feeding, care and management. Keeping animals healthy by confining purchases to healthy herds, proper quarantine at the time of bringing in new animals, employing sound principles of sanitation, management and feeding and careful use of appropriate and dependable vaccines is the practical and economical ways to avoid losses from the disease.

Good housing assists in maintaining the health of the herd, whereas judicious feeding not only builds up body resistance to disease but also helps in speedy recovery in case there is a disease attack [3]. According to Drug Administration and Control Authority of Ethiopia [4], cattle diseases are categorized into five groups, namely, infectious diseases, noninfectious diseases, diseases of the respiratory system, diseases of the reproductive system and ectoparasites.

Knowledge Base System (KBS) is one of the major family members of the Artificial Intelligence (AI) groups. With the availability of advanced computing facilities and other resources attention is now turning to more demanding tasks that might require intelligence. Societies and industries are becoming knowledge-oriented and are depending on different experts' decision making ability to solve problems [5].

The KBS is a computer program that uses artificial intelligence to solve problems within a specialized domain that requires human expertise. The typical applications of knowledge base system include classification, diagnosis, monitoring, designing, scheduling and planning for specialized tasks [6]. From these applications of knowledge base system this research work focuses on the diagnosis of disease of cattle.

Knowledge based system helps for animal health workers to upgrade their knowledge and experience and give high level medical health care services and treatment. Developing a Knowledge Base System that gives advice about cattle disease would play a critical role, so that animal health workers can save many lives. Besides of this, developing the knowledge based system encourage and enable cattle owners who have computer access to diagnosis his/her cattle to check the severity of the disease and go to the higher clinic for further diagnosis.

Data mining (DM) is a subfield of machine learning that enables finding interesting knowledge (patterns, models and relationships) from very large dataset. It is the most essential part of the knowledge-discovery process, which combines databases, statistics, and artificial intelligence machine learning techniques. Data mining improves decision making by giving insight into what is happening today and by helping predict what happened tomorrow [7].

1.2. Statement of the Problem

In Ethiopia, the number of veterinarians was very low until the last decade. The ratio of veterinarians to animals was about 1: 500,000 [8]. Even if the government is working to increase the number of veterinarians by opening veterinary colleges, the ratio of veterinarians to animals is still not satisfactory. Recently, the government has been using animal health assistant to treat animals in rural areas. These animal health assistants have only basic knowledge about animal diseases. Animal health assistants, who work in the rural town of Ethiopia, are the primary means to treat cattle in the country. There is no way to acquire information about cattle diseases after they have graduated. The information they obtain through agricultural development office is static and does not respond to users specific needs since the information is general. For example, if a new disease occurred in the country, detail symptom of the disease would not be known. During this situation, people would look for veterinarians [9]. But veterinarians are rarely able to devote adequate time to assisting all requests. Moreover, in many instances, response by veterinarians is not also on time. The use of knowledge based system could be a possible solution for treating on time. The best knowledge is no such a system in use for cattle diseases treatment, especially in Ethiopia. Thus, this research explores the applicability of using knowledge based system technology in the domain of cattle disease diagnosis and treatment by developing a knowledge based system [5].

Above all, differential diagnosis of diseases takes long time. The veterinary service cycle is delayed in the veterinarian clinics. Because there are some diseases that have two or more similar symptoms. This is difficult especially for inexperienced veterinarians. Knowledge sharing among experienced veterinarians and new ones is difficult due to lack of well experienced experts in all veterinary clinics in the developing country.

In this research, Data Mining result integrated with Knowledge Based System for Cattle Diseases Diagnosis and Treatment is present to reduce the problems, which will use to facilitate fast disease diagnosis and on time treatment by the veterinarians and educated producers of cattle. Because, several diseases show very similar conditions which need differential diagnosis and take long time to detect them.

In this regard, these studies explore and find answers to the following research questions:-

- What are the main attributes that can properly predict the type of cattle disease?
- Which classification algorithm is best to develop the prediction model that can predict the type that give to the diseased cattle which is infected by disease?
- How to acquire, model, represent and implement a data mining result knowledge based system for the diagnosis and treatment of selected disease?
- How to integrate data mining result with knowledge base system for developing the prototype?

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of this study is to apply the data mining techniques to integrate the knowledge base system, in order to increase the efficiency in diagnosis and treatment of common cattle disease.

1.3.2. Specific Objectives

To achieve the general objective the following specific objectives formulate:

- ❖ To acquire knowledge from domain experts, document analysis and data mining.
- ❖ To conduct preprocessing of Basso Animal health center cattle disease dataset based on a hybrid model.
- ❖ To explore suitable classification data mining technique and algorithm.
- ❖ To create a predictive model based on Basso Animal health center cattle disease dataset.
- ❖ To construct the Knowledge base system using rule-based knowledge representation approach.
- ❖ To evaluate the performance and user acceptance level.

1.4. Scope and Limitation of the Study

This study is, thus, bound to propose a prototype of a Knowledge Based System for the diagnosis and treatment of eight common diseases of cattle, namely blackleg, anthrax, Foot Mouse Disease (FMD), Lumpy Skin Disease (LSD), pneumonia, mastitis, enteritis and internal parasite frequently occurs, according to Basso Animal Health Center dataset. Knowledge acquisition, modeling, representation and development of the knowledge-based system for the diagnosis and treatment of selected diseases are the major goals of this research.

The time and resource constraints are limited to cover the entire KBS of all diseases in this study. Hence, this study does not include the diagnosis and treatment of all diseases of cattle. The reason is many of the viral and bacterial diseases need sophisticated laboratory equipment for the identification of them. It is, thus, limited to the differential diagnosis and treatment of diseases of cattle based on symptoms. Generally, the study tries to implement a prototype Knowledge Based

System with Java graphical user interface which plays the differential diagnosis of a limited number of diseases.

The main limitation that the researcher faced while doing this researcher is data size and quality. The dataset was small in number because of this the researcher faced a class imbalance problem and forced to use the Synthetic Minority Oversampling Technique (SMOTE) to make the imbalance dataset balanced and get as many rules and knowledge as possible.

1.5. Significance of the Study

The benefit of this study is providing the medical diagnosis and treatment of diseases of cattle. The veterinary who work in agricultural sectors can easily access the collected knowledge in the Knowledge Base System to support the cattle producers timely and quickly. Domain Experts and middle level professionals can improve their knowledge and experience using the knowledge base system to diagnose and treat diseases. KBS enables human experts to share their private knowledge among users in the agricultural fields to improve quality and productivity of cattle production.

Developing a Knowledge Base System would play a vital and critical role in documenting guidelines and knowledge and experience of well-educated and experienced veterinaries and make it available for those animal health workers, which in turn serves as means of knowledge transfer such that the people of Ethiopia could get early and timely diagnosis and treatment where ever they are which also help the country in providing a quality health care service for the people who are living in remote areas.

Developing and implementing a Knowledge Base System that provides advice to animal health workers and cattle owners would play a great role because KBS provides the high-quality performance which solves difficult problems in a domain as good as or better than human experts and can possesses vast quantities of domain specific knowledge to the minute details which can in turn serve as means of knowledge and experience transfer because the ability of the intelligent systems to capture and redistribute expertise has significant implications on development of a nation, commodity or population.

1.6. Research Methodology

1.6.1. Research Design

Since the research topic needs several experiments and domain experts' opinion, the researcher used mixed (both quantitative and qualitative) research approach. The researcher also collected the dataset for data mining purpose from Basso Animal health center cattle disease.

1.6.2. Literature Review

The researcher reviewed different researches, articles and journal papers that cattle disease and treatment, data mining, Knowledge base Systems, knowledge acquisition of knowledge based systems and integration of data mining results with knowledge base systems to get conceptual understanding about the problem on the hand.

1.6.3. Methods of Knowledge Collected

1.6.3.1. Manual Knowledge Acquisition

The researcher used both interview and document analysis to acquire knowledge. The researcher conducted the domain expert's interview with veterinaries who works for cattle diseases Treatment and Research Center at the Bassona Worena Wereda Animal Health Center.

1.6.3.2. Automated Knowledge Acquisition

The researcher used hybrid model to acquire knowledge from Basso Animal health center cattle disease using the (Waikato Environment for Knowledge Analysis **(WEKA)** data mining tool. This model consists of six phases intended as a cyclical phases, namely domain understanding, data understanding, data preparation, data mining, data evaluation and knowledge discovery

1.6.4. Methods Knowledge Representation

Since the knowledge that the researcher acquired from data mining classification technique are in the form of rules and the knowledge that the researcher acquire from document analysis and domain experts' interview about diagnosis and treatment of cattle disease are full of decision trees and procedures which are easy to convert to rules, the researcher forced to use rule-based knowledge representation method which is the most predominant knowledge representation methods to develop the Knowledge base.

1.6.5. Implementation Tool

In order to mine the hidden knowledge from the pre-processed dataset and compare the performance of classifiers, we used WEKA data mining tool. To represent rules in knowledge base and construct the prototype of cattle disease advising Knowledge Based System, the researcher use SWI-Prolog. Java NetBeans IDE 8.2 with JDK-8u20 [10] was employed to integrate WEKA result with the Knowledge based system and develop the GUI of the propose system.

1.6.6. Evaluation Methods

The researcher used True Positive rate, Precision, Recall and F-measure evaluate the results and accuracy of the developed model. The researcher also evaluate the KBS using system performance testing by preparing test cases and users' acceptance testing questionnaire which helps the researcher to make sure that whether the potential users would like to use the propose system frequently and whether the propose systems meets user requirements.

1.7. Organization of the Study

This study comprises eight chapters. Chapter one discusses background of the study, the problem statement and research questions, the objectives of the study, scope and limitation of the study and methodologies that the researcher uses to conduct this study.

Chapter two discusses about conceptual and related works review that are relevant for this study. In this chapter, the researcher discussions about data mining process model, data mining tasks, classification algorithm. In addition to this, the researcher discusses about Knowledge Based System such as methods of knowledge acquisition, knowledge representation, Knowledge Base System architecture, Knowledge Base development tools which are relevant for this study.

Chapter three presents the knowledge acquisition and modeling. In this chapter, the researcher discussions about common cattle disease, symptoms of the disease, transmission, and prevention methods, clinical diagnosis and treatment of cattle disease. The focus here is on manual (domain expert interview and documents analysis) and automated knowledge acquisition techniques through data mining. In addition to this the researcher discussed about conceptual modeling for cattle disease diagnosis and treatment.

Chapter four discusses about methodology and approach of the study. Here, the researcher presents data mining framework for mining knowledge from domain expert and the researcher presents the data mining process model domain understanding, data understanding, data understanding, data preparation, data mining, data evaluation and knowledge discovery.

Chapter five discusses about expermental result and discussion. In this chapter, the researcher discuss predictive model creation and experimentation. The researcher also discusses the results of WEKA classifier algorithms by comparing one to another and also discuss the research question.

Chapter six discusses about how to integrate data mining result with knowledge based system. In this chapter, the researcher discuss about integrated knowledge based, calling Java from Prolog, reading the ARFF Files, setting class Index then getting and classifying the instance

Chapter seven discusses about implementation and evaluation of the proposed systems. In this chapter, the researcher discuss the components of the proposed system such as the knowledge base, the inference engine, the user interface. Besides to this, the researcher discusses how to evaluate the proposed system using test cases and user acceptance testing mechanisms.

Finally, chapter eight conclusion and recommendation which is the last chapter of the study, discusses on the conclusion and forwards recommendation for further investigation.

1.8. Definition of Technical Terms

- ✓ **Classification**: - in DM, it is the task of assigning attributes to a predefined class.
- ✓ **Classification model**: - is known as a target function. It is used for predicting or describing a class.
- ✓ **Data mining** (**DM**): - process of retrieving useful information from a large amount of data.
- ✓ **Supervised learning**: - a type of DM techniques that uses training data to train the algorithm and to generate a classification model.
- ✓ **Test dataset**: - data that are used to test an algorithm for accuracy.
- ✓ **Training dataset**: - data that are used to train a supervised learning algorithm.
- ✓ **Unsupervised learning**: - a type of DM techniques that does not need to use training data to train the algorithm. No prediction involved.
- ✓ **Instance**: - an attribute or a feature of data
- ✓ **WEKA**: - is a DM software program to analyses data.
- ✓ **SWIWEKA: -** an interface that allows using the waka API for classification

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

2.1. Introduction

Cattle are large grazing animals with two-toed or cloven hooves and a four-chambered stomach. This stomach is an adaptation to help digest tough grasses. Cattle can be horned or polled (or hornless), depending on the breed. The horns come out on either side of the head above the ears and are a simple shape, usually curved upwards but sometimes down. Cattle usually stay together in groups called herds. One male, called a bull usually have a number of cows in a herd as his harem. The cows usually give birth to one calf a year, though twins are also known to be born.

The calves have long strong legs and can walk a few minutes after they are born, so they can follow the herd. Cattle are the most common type of large domesticated hoofed animals. They are a prominent modern member of the subfamily Bovine [11].

2.2. Data Mining Process Model

One of the greatest strength of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets [12]. Data mining tools perform data analysis and uncover important data patterns, contributing greatly to different business strategies including medical researchers. The widening gap between data and information calls for a systematic development of data mining tools that would turn data toms into golden nuggets of knowledge. Thus, patterns and knowledge from data mining is using for a sound judgement and proactive decision making in different organization including health care sectors. Broadly used methodologies in data mining are Knowledge Discovery in Database (KDD), CRoss Industry Standard Process for Data mining (CRISP-DM), Sample Explore Modify Model Assess (SEMMA) and hybrid process [12].

2.2.1. Knowledge Discovery in Database (KDD) Process

It is a process using data mining methods to extract what is estimated knowledge according to the specification of measures and thresholds, using the data base along with any required preprocessing, subsampling, and transformation of the data base [13]. KDD is a five stage process model, as listed in the following:

Selection: - this stage consists on creating a target dataset, or focusing on a subset of variables or data sample, on which discovery is to be performed.

Preprocessing: - this stage consists of the target data cleaning and preprocessing in order to obtain consistent data.

Transformation: - this stage consists of the transformation of the data using dimensionality reduction or transformation method.

Data mining: - this stage consists of searching for patterns of interest in a particular representational form, depending on the data mining objective.

Interpretation/Evaluation: - this stage consists of the interpretation and evaluation of the mined patterns.

The KDD process is interactive and iterative, involving numerous steps with many decisions making made by the user. Additionally, the KDD process must be presented by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end users. It also must be continued by the knowledge consolidation by incorporating this into the system [14].

2.2.2. Sample Explore Modify Model Assess Process

It is developed by the SAS institute. The acronym SEMMA stands for Sample Explore, modify, Model and Assess and refers to the process of conducting a data mining project [15].The SAS institute considers a cycle with five stages for the process lists as following:

Sample: - this stage consists on sampling the data by extracting a portion of a large dataset big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out at being optional.

Explore: - this stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify: - this stage consists on the modification of the data by creating, selecting and transforming the variables to focus the model selection process.

Model: - this stage consists on the modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcomes.

Assess: - this stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

Although the SEMMA process is independent from data mining chosen tolls, it is linked to SAS Enterprise miner software and pretend to guide the user on the implementation of data mining applications. SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of data mining project [15].

2.2.3. CRoss Industry Standard Process for Data mining Process

CRISP-DM methodology is applied to build the mining models. It consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Business understanding phase focuses on understanding the objectives and requirements from cattle disease perspective, converting this knowledge into a data mining problem definition, and designing a preliminary plan to achieve the objectives. Data understanding phase uses the raw the data and proceeds to understand the data (medical terminologies, tests etc.), identify its quality, gain preliminary insights, and detect interesting subsets to form hypotheses for hidden information. Data preparation phase constructs the final dataset that can be fed into the modeling tools. This includes table, record, and attribute selection as well as data cleaning and transformation. The modeling phase selects and applies the best models and calibrates their parameters to optimal values. The evaluation phase evaluates the model to ensure that it achieves the business objectives. The deployment phase specifies the tasks that are needed to use the models [16].

2.2.4. Hybrid Process

The development of academic and industrial model has led to the development of hybrid models, i.e., models combine aspects of both. One such model is a six step Knowledge Discovery Processes (KDP) model developed by Cios and other scholars [17]. It was developed based on the CRISP-DM model by adopting it to academic research. The main difference and extensions includes as follows:

- Providing more general, research-oriented description of the steps
- Introducing a data mining steps instead of the modeling step
- Introducing the several new explicit feedback mechanisms, (the CRISM-DM model has only three major feedback sources while the hybrid model has more detailed feedback mechanisms).
- Modification of the last steps, since in the hybrid model, the knowledge discovered for particular domain can be applied in other domain.

Summary of correspondence between KDD, SEMMA, CRISP-DM and hybrid models are presented in table 2.1:

Table 2.1 Summary of data mining models

KDD	SEMMA	CRISP-DM	Hybrid
Pre KDD	--------------------	Business understanding	Domain understanding
Selection	Sample	Data understanding	Data understanding
Preprocessing	Explore		
Transformation	Modify	Data preparation	Data preparation
Data mining	Model	Modeling	Data mining
Interpretation/ Evaluation	Asses	Evaluation	Data evaluation
Post KDD	-------------------	Deployment	Knowledge discovery

From the above table 2.1 by doing the comparisons of the models, some of them follow the same steps to discovery process whiles others follow different steps. For example in KDD and SEMMA stages the first approach is equivalent. Sample can be identified with selection. ; explore can be identified with preprocessing; modify can be identified with transformation; model can be identified with data mining; assess can be identified with interpretation/evaluation. CRISP-DM compare to hybrid data mining is most of the components are similar but on data mining identified with modeling; knowledge discovery can be identified with deployment. There for CRISP-DM is mostly used to project work but hybrid is used to research work [18]. According to this comparisons hybrid data modeling is best for this study. Data mining process, as depicted in figure 2.1 below, is a step in hybrid process which consists of methods that produce useful patterns or models from the data.

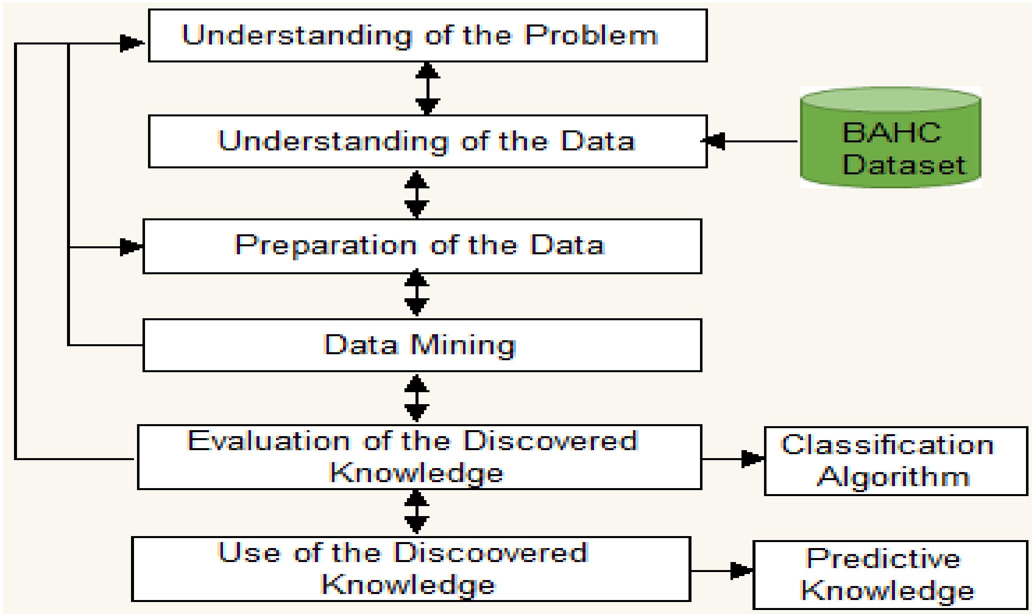


Figure 2.1 The six steps KDP Model Framework

2.3. Data Mining Tasks

Many of the problems can be phrased in terms of the following data mining tasks, which can be found in available data mining literature [19]:

2.3.1. Association Rule Discovery

It is a descriptive data mining task which includes determining patterns, or associations between elements in datasets. Associations are represented in the form of rules. The association technique is used for associating tasks [19] . Examples are market basket analysis, which more or less is determining what things go together in a shopping cart at the supermarket, and cross selling programs, which helps to design attractive packages or groupings of products and services [20].

2.3.2. Clustering

Clustering is identifying similar groups from unstructured data. Clustering is the task of grouping a set of objects in a such a way that object in same group are more similar to each other than to those in other groups. Once the clusters are decided, the objects are labelled their corresponding clusters, and common features of the objects in cluster are summarized to form a class description [21]. The difference with the classification task is that clusters were unknown at the time the algorithm starts. In other words, there are no predefined classes it relies on [20]. The records are grouped together on the basis of self-similarity. The meaning of the results of clustering are to be determined by the user. Clustering is often done as a prelude to some other data mining task. Clustering technique is used for the clustering task [19].

2.3.3. Regression

Regression, sometimes also called estimation, is a kind of statistical estimation technique which is used to map each data object to a real value provided prediction value [22]. It deals with continuously valued outcomes and comes up with a value for some unknown continuous variable. The estimation approach has the great advantage that the individual records can be rank ordered according to the estimate. Uses of regression include prediction, modelling of causal relationships, and testing hypotheses about relationships between variables. Well suited techniques for regression tasks are (linear) regression models and none linear regration [19].

2.3.4. Classification

Classification is the task of assigning class labels to the data according to a model learned from the training data where the classes are known. Classification is one of the most common tasks in supervised learning, but it has not received much attention in temporal data mining [23]. The classification task is characterized by a well-defined definition of the class labels, and a training set consisting of reclassified examples. The task is to develop a classification model of some kind that can be applied to unclassified data in order to classify it. Decision trees, nearest neighbor are well-suited techniques for the classification task.

The developed model is based on the analysis of a set of training data whose class label is known and the derived model may be represented in various forms such as IF-THEN rules, decision trees, mathematical formulae, semantic network etc. [24]. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attributes set and the class label of the input data. The following classification model illustrates the data mining algorithm implementation shows in Figure 2.2 as follows:

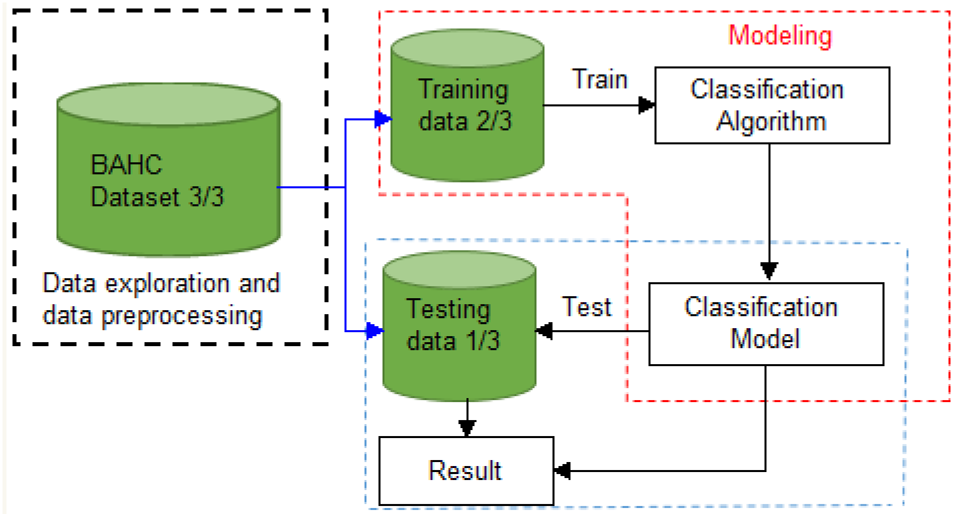


Figure 2.2 Data Mining Algorithm Implementation Classification Model *[25]*

There are various classification techniques; among which the main classification techniques are lists as the following [24].

2.3.4.1. Decision Tree Induction

The decision tree algorithm is probably the most popular data mining technique because of the fast training, performance, a high degree of accuracy, and easily understandable patterns. Splitting your data into subsets is the main idea behind the algorithm [26]. When decision tree induction is used for attributes subset selection, a tree is constructed from the given labeled data. All attributes that do not appear in the tree are assumed to be irrelevant. There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literatures that construct decision trees from a set of input-output training samples [27]. In decision tree construction, selection of splitting attributes is necessary in order to avoid irrelevant attributes by examining the effect of each attribute for the distinct class and its likelihood for improving the overall decision performance of the tree, since the feature with minimum impact on dependent variable may distort the tree's performance and the classification accuracy.

One of the most attractive aspects of decision trees lies in their interpretability especially with respect to the construction of decision rules which is constructed from a decision tree simply by traversing any given path from the root node to any leaf [27]. Therefore, to make a decision tree model more readable, a path to each leaf can be transformed into an IF-THEN rule. In Figure 2.3, illustrates the root node and leaf node as follows [28].

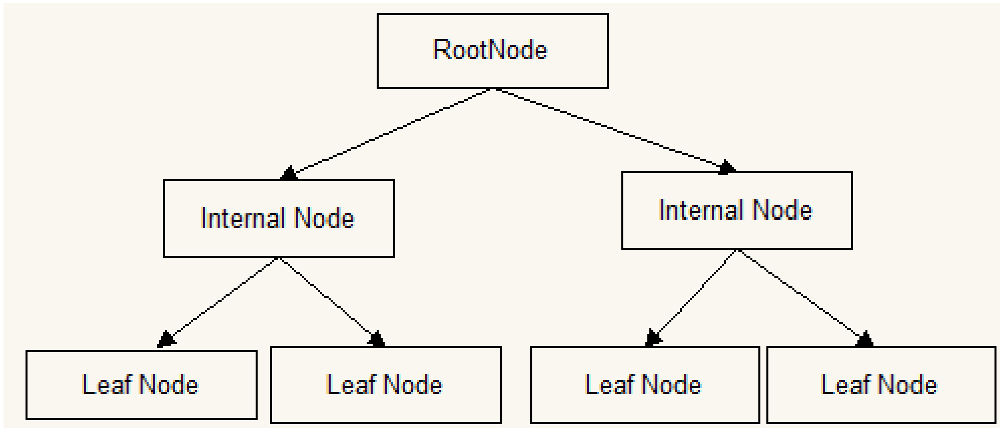


Figure 2.3 Decision tree Structure

The challenge with decision tree is overfitting. As the dataset grows larger and the number of attributes grows larger, we can create trees that become increasingly complex [24]. This potentially leads to the concept of overfitting which consequently brings the notion of pruning; this implies removing of the branches of the classification tree in order to make tree as simple and compact as possible, with as few nodes and leaves as possible. This is done through pruning a tree by halting its construction by partition the subset of training tuples at a given node or removing sub trees from a fully grown tree [24].

The main advantages of decision trees over other algorithms are that they are quick to build, efficient and easy to understand as each node is labelled in terms of the input attributes. The basic algorithm for decision tree induction is greedy algorithm that constructs decision trees in a top-down recursive divide and conquer manner [29]. The algorithm is summarized as follows.

Create a node N;

If samples are all of the same class, C then

Return N as a leaf node labeled with the class C;

If attribute-list is empty then

Return N as a leaf node labeled with the most common class in samples;

select test-attribute, the attribute among attributes-list with the highest information gain; label node N with test-attribute; for each known value AI of test-attribute grow a branch from node N for the condition test-attribute= AI;

let s_i be the set of samples for which test-attribute= a_i;

If s_i is empty then attach a leaf labeled with the most common class in samples;

else attach the node returned by

Generate_decision_tree(si,attribute-list_test-attribute)

J48 Classification Algorithm

Decision tree J48 implements Quinlan's C4.5 algorithm for generating a pruned or un-pruned C4.5 tree. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets.

2.3.4.2. Rule Based Classification

Though the decision tree is a widely used technique for classification purposes, another popular alternative to decision trees is classification rules which can be expressed as parts IF-THEN rules so that humans can understand them easily [27]. It is a relationship between antecedent, and consequent i.e. an expression of the form IF condition THEN the conclusion. The algorithm decision tree is the best known method for deriving rules from classification trees [27].

The advantage of IF-THEN rule is the rules are order independent, i.e. regardless of the order of rules executed, the same classification of the classes is possible to reach [27]. The challenges are the generated rules are often more complex than necessary and contain redundant information and the rules generated this way may be unnecessarily complex and incomprehensible [27].

Rule based classifier group instances by using a set of IF… THEN rules.

Rule :(Condition) → X

Where,

- **Condition** is a conjunction of attributes like (A_1=v_1) and (A_2=v_2) and …and (A_n=v_n) and
- **X** is a class label.

For example: (fever=high) ^ (diarrhea>= Yes) → Enteritis

JRip

JRIP is a prepositional rule learner, i.e. Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [30]. Rules in this algorithm are generated for every class in the training set and are then pruned. The discovered knowledge is represented in the form of IF-THEN prediction rules, which have the advantage of being a high-level and symbolic knowledge representation contributing towards the comprehensibility of the discovered knowledge.

JRip is based on the construction of a rule set in which all positive instances are covered by partitioning the current set of training instances into two subsets namely a growing set and a pruning set. The rule is constructed from instances in the growing set. Initially, the rule set is empty and the rules are added incrementally to the rule set until no symptom instances are covered. Following this the algorithm substitutes or revises individual rules by using reduced error pruning in order to increase the accuracy of the rules. To prune a rule the algorithm takes into account only a final sequence of conditions of the rule and sorts the deletion that maximizes the function [31].

2.3.4.3. Naïve Bayes Classifier

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given dataset. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naïve Bayes yet the algorithm tends to perform well and learn rapidly in various supervised classification problems [32] . Steps to calculate Naïve Bayes formula as follows:

Step 1: Convert the dataset into a frequency table

Step 2: Create Likelihood table by finding the probabilities

Step 3: Use Naïve Bayesian formula to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

$$\text{Posterior probability P (C|X)} = \frac{Likekihood\ P(X|C) * Class\ prior\ Probabilit\ P(c)}{\text{Predictor Prior Probability P(x)}} \quad \ldots\ldots. (2.9)$$

2.4. Attribute Selection Measures

An attribute selection measure for developing decision tree is a heuristic for selecting the splitting criterion that best separates a given data partition of class-labeled training instances into individual classes. The attribute selection measure provides a ranking for each attribute describing the given training instances. The attribute that has the best score for the measure is chosen as the splitting attribute for the given instance. The tree node created for partition, let‘s say D, is labeled with the splitting criterion, branches are grown for each outcome of the criterion, and the instances are partitioned accordingly [24]. This section describes three popular attribute selection measures, namely information gain, gain ratio, and Gini index.

2.5.1. Information Gain

Information gain for attribute selection measure is based on the work of Claude Shannon on information theory, which studied the value or information content of messages. **Iterative Dichotomiser 3** (ID3) uses information gain for attribute selection measure. The notion used is as follows: - Let D, the data partition, be a training set of class labeled instances. Suppose the class label attribute has m distinct values defining m distinct classes, C_i(for i=1, …, m). The attribute with the highest information gain is selected as the splitting attribute. This attribute minimizes the information needed to classify the instances in the resulting partitions and reflects the least impurity in these partitions. Entropy (impurity) is used to measure the information content of the attributes. High entropy means the attribute is from a uniform distribution whereas low entropy means the attribute is from a varied distribution. Entropy is defined as follows. Let p_i be the probability that an arbitrary instance, in D belongs to class $C_{i,}$ estimated by $|C_{i,\ D}|/|D|$. Expected information (entropy) needed to classify an instance in D is given shows in the following equation 2.1:

$$\text{Entropy (E(D))}=\sum_{i=1}^{m} pi\ log(pi)\ldots\ldots\ldots\ldots\text{(2.1)}$$

Entropy(E(D)) - is the average amount of information needed to identify the class label of an instance in D. The smaller information required, the greater the cleanliness. At this point, the information we have is based solely on the proportions of instances of each class. A log function to the base 2 is used, because the information is encoded or measured in bits.

Suppose attribute A can be used to split D into v partitions or subsets, {D1,D2,…, Dv}, where Dj contains those instances in D that have outcome aj of A. Information needed (after using A to split D) to classify D:

$$\mathrm{Info}_A(D)=\sum_{j=1}^{v}\frac{Dj}{D}\ \mathrm{X}\ \ \mathrm{Info}(Dj)............(2.2)$$

The smaller the expected information required, the greater the purity of the partitions. Information gained by branching on attribute A is given by

$$\mathrm{Gain}(A)=E(D)-\mathrm{Info}_A(D)……..(2.3)$$

Information gain increases with the average purity of the subsets. The attribute that has the highest information gain among the attributes is selected as the splitting attribute.

2.5.2. Gain Ratio

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. This may result in the selection of an attribute that is non-optimal for prediction. C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a split information‖ value defined analogously with Info(D) as:

$$\mathrm{SplitInfo}_A(D)= -\sum_{j=1}^{v}\frac{|Dj|}{|D|}\ \mathrm{X}log\frac{|Dj|}{|D|}………(2.4)$$

This value represents the potential information generated by splitting the training dataset, D, into v partitions, corresponding to the v outcomes of a test on attribute A. Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D. It differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning [24]. The gain ratio is defined as:

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

The attribute with the maximum gain ratio is selected as the splitting attribute. Note, however, that as the split information approaches 0, the ratio becomes unstable. A constraint is added to avoid this, whereby the information gain of the test selected must be large at least as great as the average gain over all tests examined.

2.5.3. Gini Index

The Gini index is used in Classification & Regression Trees CART. Using the notation described above, the Gini index measures the impurity of D, a data partition or set of training tuples [7], as described in eq. 2.6 as follows:

$$\text{Gini}(D)=1-\sum_{i=1}^{m} pi^2 \ldots\ldots..(2.6)$$

Where pi is the probability that a tuple in D belongs to class Ci and is estimated by $|C_{i,D}|/|D|$. The sum is computed over m classes. To determine the best binary split on A, we examine all the possible subsets that can be formed using known values of A and need to enumerate all the possible splitting points for each attribute. If A is discrete valued attribute having v distinct values, then there are 2v-v possible subsets. When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. If dataset D is split on A into two subsets D1 and D2, the Gini index gini(D) is defined as [24]:

$$\text{Gini}_A(D)=\frac{|D1|}{|D|}\text{Gini}(_{D1)}+\frac{|D2|}{|D|}\text{Gini}(_{D2})\ldots\ldots..(2.7)$$

First, we calculate Gini index for all subsets of an attribute, then the subset that gives the minimum Gini index for that attribute is selected. The point, giving the minimum Gini index for a given (continuous valued) attribute is taken as the split-point of that attribute. The reduction in impurity that would be incurred by a binary split on attribute A is

$$\Delta Gini(A) = \text{Gini}(D) - \text{Gini}_A(D)\ldots\ldots..(2.8)$$

The attribute that maximizes the reduction in impurity (or has the minimum Gini index) is selected as the splitting attribute.

To summarize, the three measures for attribute selection are used mostly. Information gain is biased towards multi valued attributes. Whereas Gain ratio tends to prefer unbalanced splits in which one partition is much smaller than the others. Gini index is biased to multi valued attributes and has difficulty when the number of classes is large. The algorithm used each attribute of the data to make decisions by splitting the data into smaller subsets. All the possible tests are considered during decision making based on the information gain value of each attribute.

2.5. Knowledge Base System

Knowledge based systems is a branch of artificial intelligence, which is a computer program that attempts to replicate the reasoning processes of a human expert and it can make decisions and recommendations and perform tasks based on user input. Artificial Intelligence is all about how making the system think, or act like humans. The expert`s knowledge is available when the human expert might not be and so that the knowledge can be available at all times and in many places, as necessary. Expert systems derive their input for decision making from prompts at the user interface, or from data files stored on the computer [33].

Knowledge-based systems are computer programs designed to solve problems, generate new information (such as a diagnosis), or provide advice, using a knowledge base and an inference mechanism. Most systems include a user interface and some explanation capability as well. Knowledge-based systems are characterized as focusing on the accumulation, representation, and use of knowledge specific to a particular task, but addressed the expanded views of such systems made possible by the ability to use the same knowledge in several different ways [34].

There are two main parts to any Knowledge Based System (KBS) - the knowledge base and the inference engine. In addition, there are peripheral features designed to facilitate interaction with end-users (user interface), explanation of a line of reasoning [35]. Figure 2.4 illustrates the structure design of the general architecture of Knowledge Based System.

The most important ingredient in any expert system is the knowledge. The power of domain expert system resides in the specific, high-quality knowledge it contains about task domains [36] [37]. Knowledge can be classified in many different ways. Tacit knowledge, explicit knowledge, hidden knowledge, factual knowledge, procedural knowledge, commonsense knowledge, domain knowledge, Meta knowledge, etc. [38].

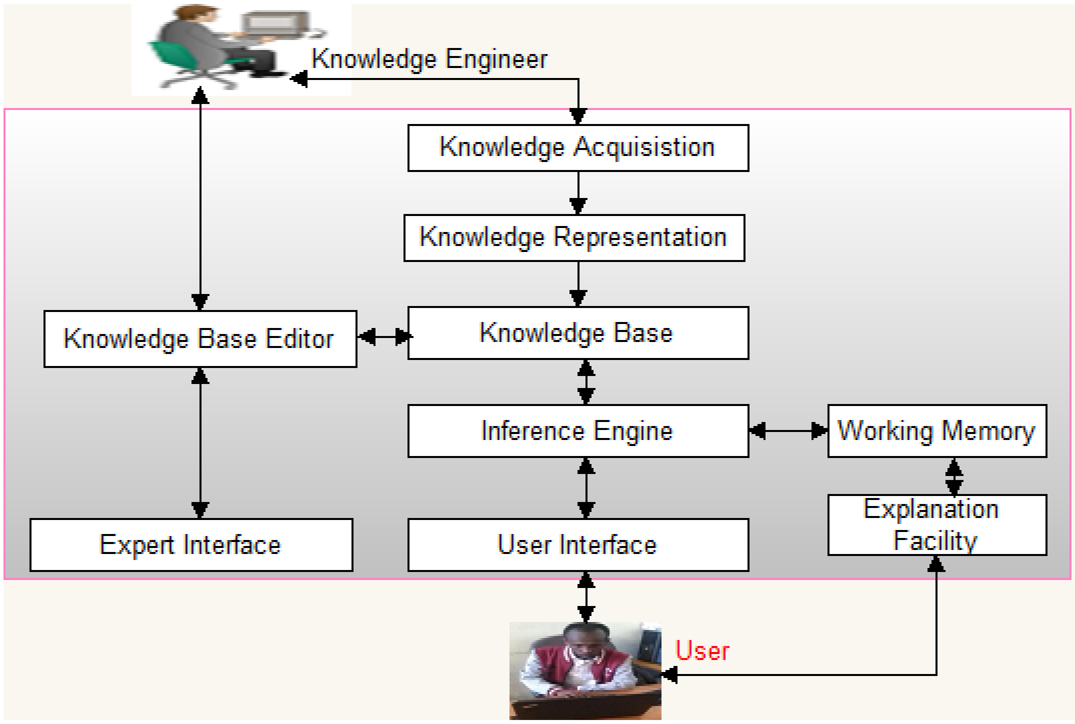


Figure 2.4 The General Architecture of Knowledge Based System *[7]*

Knowledge base: contains the knowledge necessary for understanding, formulating and for solving problems. It is a warehouse of the domain specific knowledge captured from the human expert via the knowledge acquisition module. To represent the knowledge production rules, frames, logic, semantic net, etc. is used [36]. The knowledge base stores all relevant information, data, rules, cases, and relationships used by the expert system. A knowledge base can combine the knowledge of multiple human experts [39].

Inference Engine: is a brain of expert systems. It uses the control structure (rule interpreter) and provides a methodology for reasoning. It acts as an interpreter which analyzes and processes the rules. It is used to perform the task of matching antecedents from the responses given by the users and firing rules. The major task of inference engine is to trace its way through a forest of rules to arrive at a conclusion [39].

The purpose of the inference engine is to seek information and relationships from the knowledge base and to provide answers, predictions, and suggestions in the way a human expert would. The inference engine must find the right facts, interpretations, and rules and assemble them correctly [39]. Two types of inference methods are commonly used, namely backward chaining and forward chaining. Backward chaining is the process of starting with conclusions and working backward to the supporting facts. Forward chaining starts with the facts and works forward to the conclusions [39].

User Interface: - is the interaction point between the user and the system. The user interface can be graphical user interface (GUI) or command line interface (CLI). In the course of integration of data mining with knowledge based system, Graphical User Interface for the integrator and Command Line Interface for the knowledge based system.

Explanation facility: - it provides information to user for the questions asked by the system. This facility is helpful to have clarity while answering to the questions asked by the system. In steps identifying disease as blackleg, anthrax, foot mouse disease, pneumonia, lumpy skin disease, mastitis, enteritis, and internal parasite the system displays questions. The user can get explanation of the question asked if he/she has no clarity with it.

Knowledge Acquisition: - It is the accumulation, transfer and transformation of problem-solving expertise from experts and/or documented knowledge sources to a computer program for constructing or expanding the knowledge base. It is a subsystem which helps experts to build knowledge bases [36].

The knowledge acquisition process incorporates typical fact finding methods like interviews, questionnaires, record reviews and observation to acquire facts and explicit knowledge. The knowledge acquisition process is usually comprised of three principal stages [7]:

- Knowledge elicitation is the interaction between the expert and the knowledge engineer/program to elicit the expert knowledge in some systematic way.
- The knowledge thus obtained is usually stored in some form of human friendly intermediate representation.
- The intermediate representation of the knowledge is then compiled into an executable form (e.g. Production rules) that the inference engine can process.

This is a time intensive process, and automated knowledge elicitation and machine learning techniques are increasingly common modern alternatives stages of knowledge acquisition [7].

2.6. Knowledge Representation

Knowledge Representation (KR) is the area of artificial intelligence concerned with how knowledge can be represented symbolically and manipulated in an automated way by reasoning program. Knowledge is a progression from data to information, from information to knowledge and from knowledge to wisdom while representation is a combination of syntax, semantics and reasoning. Hence, KR is an area of research in AI which aims at representing knowledge in symbols to facilitate inference from those knowledge elements, creating new elements of knowledge [40].

As a branch of symbolic Artificial Intelligence, knowledge representation and reasoning aims at designing computer systems that reason about a machine-interpretable representation of the world, similar to human reasoning. In Artificial Intelligence, knowledge representation studies the formalization of knowledge and it's processing within machines [41]. The object of knowledge representation is to express knowledge in computer-tractable form, such that it can be used to help agents perform well [42].

2.6.1. Forms of Representing Knowledge

2.6.1.1. Semantic Networks Knowledge Representation

A semantic network is a graph whose nodes represent concepts and whose arcs represent relations between these concepts. They provide a structural representation of statements about a domain of interest. Semantic networks provide a means to abstract from natural language, representing the knowledge that is captured in text form more suitable for computation [41].

Semantic networks are closely related to another form of knowledge representation called frames systems. In fact, frame systems and semantic networks can be identical in their expressiveness, but use different representation metaphors. While the semantic network metaphor is that of a graph with concept nodes linked by relation arcs, the frame metaphor draws concepts as boxes, i.e. frames, and relations as slots inside frames that can be filled by other frames. Thus, in the frame metaphor the graph turns into nested boxes [41].

The semantic network form of knowledge representation is especially suitable for capturing the taxonomic structure of categories of domain objects and for expressing general statements about the domain of interest. Inheritance and other relations between such categories can be represented in and derived from subsumption hierarchies. On the other hand, the representation of concrete individuals or even data values, like numbers or strings, does not fit well the idea of semantic networks [41].

2.6.1.2. Rules Based Knowledge Representation

It is important to emphasize that logic is not the knowledge itself; it is simply a way of representing knowledge. However, logic can be viewed as a form of Meta level knowledge about how to represent and reason with knowledge. What logic enables us to do is represent the knowledge possessed by an agent using a finite set of logical expressions plus a process (namely, the inference rules of logic) for generating a (potentially unlimited) set of other logical expressions that are part of the agent's knowledge [43]. Rules come in the form of IF-THEN-constructs and allow expressing various kinds of complex statements. Rules can be found in logic programming systems, like the language Prolog, in deductive databases or in business rules systems. The IF-part is also called the body of a rule, while the THEN-part is also called its head. Typically, rule-based knowledge representation systems operate on facts, which are often formalized as a special kind of rule with an empty body. They start from a given set of facts and then apply rules in order to derive new facts, thus "drawing conclusions" [41].

It is important to distinguish between facts and their representations. Facts are part of the world, whereas their representations must be coded in some way that can be physically stored within an agent. We cannot put the world inside a computer (nor can we put it in a human), so all reasoning mechanisms must operate on representations of facts, rather than on the facts themselves. Because, the sentences are physical configurations of parts of the agent, reasoning must be a process of constructing new physical configurations from old ones. Proper reasoning should ensure that the new configurations represent facts that actually follow from the facts that the old configurations represent [42].

Among the various methods of knowledge representation methods rule based is the most commonly used forms of knowledge representation for this study and they are derived from popular techniques such as decision trees.

2.6.2. Implementation Tools

In order to mine hidden knowledge from the pre-processed dataset and compare the performance of classifiers, WEKA 6.8.1 is used. WEKA is chosen since it is proven to be powerful for data mining and used by many researchers for mining task and the researcher is familiar with the tool. It contains tools for data preprocessing, clustering, regression, classification, association rules and visualization. WEKA is written in the Java language and contains a GUI for interacting with data files and producing visual results.

In addition, in order to develop an application which maps the knowledge acquired from the data mining classifiers with knowledge based system Java NetBeans IDE 8.2 with JDK -8u20 is employed. NetBeans offers easy and efficient project management, has better support for latest Java technologies, and can be installed on all operating systems supporting Java. To represent rules in knowledge base and constructing the Rule based diagnosis disease and advising Knowledge based system PROLOG is used.

2.7. Evaluation of the Classification Model

In order to evaluate the performance of the classifier Prediction Accuracy, True Positive, False Positive, Precision, Recall and F-Measure are commonly used. Confusion matrix helps to see a breakdown of a classifier‘s performance by showing how frequently instances of a class let us say class X are classified as class X or misclassified as some other class, say class Y [44].

Table 2.3 Confusion Matrix

		Predicted class		Total instances
		+	-	
Actual Class	+	True Positive	False Positive	Positive
	-	False Negative	True Negative	Negative

2.7.1. Prediction Accuracy

Prediction accuracy measures the proportion of instances that are correctly classified by the classifier.

Predictive Accuracy = $\frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}}$ ------------------ (2.9)

Where, TP is true positive rate, TN is true negative rate, FP is false positive rate and FN is false negative rate.

2.7.2. True Positive Rate and False Positive Rate

In contrary to Predictive Accuracy TP rate and FP rate values do not depend on the relative size of positive and negative classes [44].

True Positive rate (TP) is the proportion of positive or correctly classified instances as positive or correct instances.

True Positive rate= $\frac{TP}{TP+FN}$ ---------------------------- (2.11)

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

False Positive rate= $\frac{\mathbf{TN}}{\mathbf{TN+FP}}$ ------------------------------ (2.12)

2.7.3. Precision and Recall

Precision: is measuring the proportion of instances that are classified as positive that are really positive.

Precision = $\frac{TP}{TP+FP}$ ---------------------------------- (2.13)

Recall: is what percent of positive/ negative tuples the classifier labeled as positive or negative for both True and False Classes (Blackleg, Anthrax, FMD, LSD, Mastitis, Enteritis, Pneumonia and internal Parasite).

Recall= $\frac{TP}{TP+FN}$-------------------------------------- (2.14)

F-Measure: is calculated as the harmonic mean of recall and precision.

F-measure= $\frac{2*Precision*Recall}{precision+Recall}$ -------------------- (2.15)

2.8. Related Work

Many studies have attempt to diagnosis of cattle disease in order to diagnosis and treatment of process using different techniques including expert system, machine learning and Knowledge base system. This section investigates related works done on diagnosis and treatment of cattle disease.

Derejaw [9], has developed on Web-based expert system for diagnosing cattle diseases. On his research, Knowledge is obtained from domain experts (veterinarians). For this research, the rule based knowledge representation method is chosen because it clearly demonstrates the domain knowledge. There are already defined set of symptoms, syndromes and basic issues that should be addressed to confirm the presence of cattle diseases. As a result, rule based representation method is more appropriate to represent and demonstrate the real domain knowledge in diagnosing cattle diseases.

The researcher evaluates the performance of the result is by gathering information from the domain expert. Therefore, for the comfort of analyzing the relative performance of the system based on user evaluation to evaluate the performance. The overall average performance accuracy of the system is about 87.2%. This implies that the modeled system performs well in making right decisions in the diagnosis of cattle diseases. Generally, we can conclude that the respondents are satisfied with the following criteria: user interface design, accuracy, response time and significance of the system [9].

Ahmad et al [45]**,** conducted the research on web based cattle disease expert system diagnosis with forward chaining method in Indonesia. The researcher have used forward chaining and certainty factor method. Production rules are one of the many forms of knowledge representation used in the development of expert systems. Representation of knowledge with the rules of production, basically a rule (rule) in the form of IF-THEN. There are two stages of the model that are often used to calculate the level of confidence (CF), of a rule, namely digging out the results of interviews with experts and Manual calculation of CF Value. Finally, the modeled system performs not well known in making right decisions in the diagnosis of cattle diseases.

Oluwatoyin and Adedoyin [46]**,** have developed a fuzzy expert system for diagnosis of cattle disease in Nigeria. In their research, they have developed a three layered architecture, namely application layer, business layer and storage layer. The system has given great advantages for the veterinary doctors. It gives reliable, high speed and accurate observations in diagnosing of cattle diseases. It can also be used in the absence of clinicians to diagnose diseases.

The system uses the Bayesian Belief Network methodology described. The system is parameterized using data collected from veterinary experts. A fuzzy rule base contains a set of fuzzy rules in a single IF-THEN rule in the form of membership value. There is a procedure for the fuzzy algorithm to diagnose the disease.

First to display the list of all the symptoms from the database. After getting the database to select the particular symptom of concern to be diagnosed and to select the level of severity of each selected symptom. Each selected symptom, search the database for all the diseases that has the symptom and then the weight of the symptom and multiply it by the appropriate linguistic variable.

If select the disease with the highest value after summation has been carried out in step four and display its corresponding disease. This is the diagnosed disease. Finally, the diagnosed disease is displayed in the application. Finally, the researchers have recommended that the work can be extended using the combined features of neural network and fuzzy logic to improving disease diagnosis in cattle.

Rong and Daoliang [47], have developed a web based expert system for diagnosis of cow disease. Their system has been composed of three-tier web application which uses Internet Information Server (IIS), Microsoft SQL server 2000, Windows XP as the operating system and Windows XP with Internet Explorer (IE) on the client side. Their proposed system have adopted three algorithm, i.e. Case Based Reasoning (CBR), subjective Bayesian theory and Dempster-Shafer (D-S) evidential theory. Their developed system consists of four subsystem. These are case management, diagnosis, prevention and cure, and medication management. To diagnose the milch cow disease, the system first look for a disease that can explain all the symptoms. If there is no such disease, it looked for a disease that can explain all but one symptom, etc., until it finds a possible disease.

Birhane [40], has developed a prototype KBS for diagnosis and treatment of diseases of sheep and goats. Common Knowledge Acquisition and Documentation Structuring (CommonKADS) knowledge engineering methodology is used to model the requirements of the system. Production rule is used to develop the prototype system as knowledge representation technique.

For this research, the rule based knowledge representation method is chosen because it clearly demonstrates the domain knowledge. There are already defined set of symptoms, syndromes and basic issues that should be addressed to confirm the presence of sheep and goat diseases. As a result, rule based representation method is more appropriate to represent and demonstrate the real domain knowledge in diagnosing sheep and goat diseases.

The researcher evaluates the performance of the result is by gathering information from the domain expert. Therefore, for the comfort of analyzing the relative performance of the system based on user evaluation to evaluate the performance. The overall average performance accuracy of the system is about 84.8%. This implies that the modeled system performs well in making right decisions in the diagnosis and treatment of sheep and goat diseases [40].

Tesfamariam [48], has developed diagnosis and treatment of system for Visceral Leishmaniosis (VL) using integration of data mining and knowledge base system. The researcher has used

Knowledge Discovery database (KDD) model to extract data mining from Médecins Sans Frontières (MSF) Holland Andurafi project dataset.

The performance of proposed system has been evaluated with the metrics of precision, recall, F-measure and positive rate as shown the in table 2.4 as follows:

Table 2.4 Performance evaluation Result

Metrics	J48	JRip	PART
TP rate	0.68	0.68	0.67
Precision	0.74	0.73	0.73
Recall	0.68	0.67	0.68
F-measure	0.65	0.65	0.65

The researcher decides to use rules that are generated by the PART classification algorithm model for further use in the development of knowledge base of the knowledge based system. Finally [48], result study archives promising result with 95% and 86% of the system performance and user acceptance respectively.

Away from each other the above researchers and as to the review of the researcher, no study was done at the area throughout the Amhara Region animal health center using data mining techniques. The researcher see that the limitation of the existing system is not considering analysis of data in detail. Hence, the aim of this study to apply data mining techniques integrate with knowledge based system in order to diagnosis of cattle disease and treatment. And also determine the existent and significant of it, and also it can be a foot step for further findings.

CHAPTER THREE

KNOWLEDGE ACQUISITION AND MODELING

3.1. Introduction

Knowledge acquisition (KA) is the process of eliciting, structuring and organizing knowledge from human experts, books, documents, and sensors. On the other hand, it refers computer files and transferring to the knowledge base using knowledge representation techniques used in KBS; namely, logic, production rules, semantic nets, and frames [49]. The step of knowledge acquisition is one of the major bottlenecks in the stage of knowledge base development. Usually, for each application domain there are several sources of knowledge (human experts, the specialized literature which includes textbooks, books, reviews, collection of data during the run of similar systems, etc.) [38].

3.2. Knowledge Acquired Using Documents Analysis

Document analysis involves gathering knowledge from existing documentations. Hence, document analysis has been carried out to acquire explicit knowledge which is found in various secondary sources of knowledge. According to Drug Administration and Control Authority of Ethiopia [4], cattle disease categorized into five groups namely, infectious diseases, noninfectious diseases, diseases of the respiratory system, diseases of the reproductive system and ectoparasites. Infectious disease is subdivided in to six namely, blackleg, anthrax, foot mouse disease (FMD), lumpy skin disease (LSD), enteritis, and internal parasite. Mastitis disease is one division of reproductive system disease and pneumonia is a respiratory system disease shows in the following Figure 3.1.

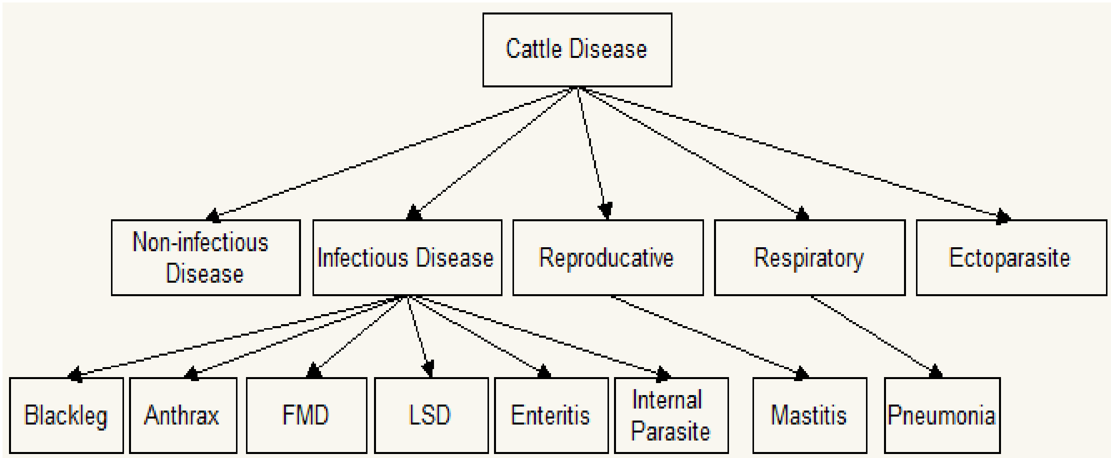


Figure 3.1 Major categories of cattle disease and its some of disease

3.2.1. Common Cattle Disease

Based on the above figure 3.1 in Basso Animal Health Center dataset common cattle disease is Foot-mouse Disease (FMD), Lumpy Skin Disease (LSD), Blackleg, Anthrax, Internal Parasite, Enteritis from infectious disease, Mastitis from reproductive system disease and pneumonia from the respiratory system disease are selected for this study.

3.2.1.1. Blackleg

Blackleg is an acute, febrile disease of cattle and sheep caused by Clostridium chauvoei characterized by emphysematous swelling of the heavy muscle and severe toxemia. Blackleg is not transmitted directly from sick animals to healthy animals by mere contact. Lameness, loss of appetite, rapid breathing, depression and high fever are the signs observed in animals infected by this disease. The sick animal usually dies within 12 to 48 hours. In most cases the animal is found dead without being previously observed sick. Blackleg is mostly occurring in the northern part of Ethiopia [50]. Cattle within the age of six months to two years old are mainly affected by this disease, though at any age and condition is affected. Blackleg is common in Ethiopia during dry period of the year [4].

Clinical symptoms

Depression, anorexia, rumen stasis, high-fever (41-42^{0}C) and painful. The marked lameness with muscle swelling of the upper part of the affected leg with capitation may follow. At necropsy affected tissues are filled with rancid serosanguineous fluid and gas pockets, which crepitate when squeezed and the muscle appear dry [4].

Diagnosis

The clinical signs and postmortem findings are indicative; the epidemiology and bacterial isolation are confirmatory [4].

Treatment and Prevention

Non-drug treatment

- ✓ Drainage and slashing of affected tissues to allow oxygen into the tissue, plus supportive treatment with parental fluids, analgesics, etc.

Drug treatment

- ✓ Procaine penicillin G, 22,000IU/kg, IM or SC q 24 h for 3 to 5 days or Benzathine penicillin or similar repository preparations, q 48-72 h.
- S/E, C/I and D/I: Hypersensitivity reactions to penicillin's, simultaneous administration of chloramphenicol, tetracycline or phenylbutazone.
- D/F: 200,000 IU/ml to 400000 IU/ml
- W/P: meat 14 days and milk 3 days
- ✓ Local antibiotic treatment Aug. Oxyteracycline spray 5% of the site of the wound is helpful.

Prevention

- ✓ Vaccination with C. chauvoeibacterin.

3.2.1.2. Anthrax

Anthrax, which is caused by bacterium, occurs in areas where animals have previously died of anthrax. Even though anthrax has appropriate vaccination, in Ethiopia, still it occurs frequently [50]. One of the common signs in cattle with anthrax is a progression from a normal appearance to dead in a matter of hours. Weakness, fever, excitement followed by depression, difficulty in breathing, uncoordinated movements and convulsions are other signs of anthrax additionally, after death, the animal's body rapidly decomposes.

Anthrax is an acute, febrile (24^0C), septicemia, fatal bacterial disease of food animals caused by Bacillus anthracis. An anthrax outbreak occurs regularly and is commonly associated with neural or alkaline, calcareous soils where the spores revert to the vegetative form and multiply to infectious levels if environmental conditions of soil, moisture, temperature, and nutrition are optimizations [4].

Clinical symptoms

The clinical sign in anthrax disease includes in ruminant species, acute illness is characterized by abrupt onset of fever, signs of abdominal pain, abortion, diarrhea, and milk production is decreasing. Chronic infection is rare in cattle and is manifested by localized edematous swelling on the ventral neck, Thorax and shoulders [4].

Diagnosis

Anthrax is diagnosed by examining blood (or other tissues) for the presence of the bacteria. Samples must be collected carefully to avoid contamination of the environment and to prevent human exposure to the bacteria [4].

Treatment and prevention

Supportive therapy

Hyperimmune serum plus antibiotics

- ✓ Penicillin 22,000 IU/kg, IM, q 12 h for 2 days, then daily for 3 days or Benzathine penicillin or other repository preparations, q 48-72 h; the initial dose should be administered IV. For C/I/, S/E, D/I, D/F, W/P, see page 34
- ✓ Oxyteracycline 6-11 mg/kg, IM, or IV, q 12-24 h. Initially, divided the daily dose into two doses.
- S/E, C/I, D/I: renal impairment, last 2-3 weeks of ingestion in pregnant animals and up to 4 weeks of age in neonates. Gastrointestinal symptoms are more severe with Oxyteracycline among the tetracycline's; discoloration of the teeth when used during pregnancy and drug interactions with anti-acids, dairy products, calcium salts, iron salts, magnesium salts, zinc salts and warfarin.
  - D/F: injection, 5, 10%
  - W/P: meat 21 days, milk 7 days
- ✓ Amoxicillin 5-10 mg/kg q 24 h for 3-5 days; for C/I, S/E, D/F, D/I, and W/P are similar to penicillin

Vaccination:

Caution: animals that have died of anthrax should be burned in a closed incinerator. Animals should not be vaccinated within 2 months of anticipated slaughter; antibiotics should not be administered with one week of vaccination.

3.2.1.3. Mastitis

It is, a complex and costly disease of dairy cows, that results from the interaction of the cow and environment including milking machine and microorganism [51]. In Ethiopia, even though the disease of mastitis has been known locally, it has not been studied systematically, making information available on the prevalence of disease and associated economic loss inadequate [52]. Unfortunately, mastitis is not always easy to detect in its early stages, particularly when the redness and swelling of the udder is not obvious. If left untreated, severe clinical mastitis may cause the death of the cow [53].

Mastitis, an inflammation of the mammary gland, is almost always due to infection by bacteria of mycotic pathogens. Although over 135 microorganisms have been reported because the disease, Staphylococcus aureus, Staphylococcus agalactiae, Str. uberis, Str. dysgalactiae, other treptococci, Arcanobacterium pyogenes, Mycoplasma spp, Nocardia asteroides and Coliforms are the most agents in Mastitis disease [4].

Clinical symptoms

Clinical mastitis is manifested by inflammation of the udder and often accompanied by abnormal milk secretions. The signs depend on the organism involved. Systemic signs could also be observed. Subclinical mastitis is the most common [4].

The clinical sign is usually non-specific, but the following gives the clue for the type of agent involved.

Table 3.1 Mastitis basic microorganism

Microorganism	Clinical finding
Staphylococcus aureus	Sever swelling, purulent milk with clots
Mycoplasma species	Drop in milk production, infection of all quarters simultaneously
Arcanobacterium pyogenes	Profuse, foul-swelling, purulent discharge
Mycoplasma bovis	Rapid onset

Diagnosis: - It is based on clinical signs, and identification of the pathogen. A test to detect subclinical mastitis include California Mastitis Test, or direct somatic cell count [4].

Treatment and Prevention

Drug Treatment:

Treatment depends on the type of infecting organism and the stage of mammary gland damage.

- ✓ Subclinical infections:

Intramammary infusion, q 48 h for 3 times, applied separately into every quarter parachute or acute staphylococcal mastitis:

- ✓ Systematic treatment

Procaine penicillin G 22, 000 IU/kg, aqueous suspension, IM or SC q 24 h for 3 to 5 days or Benzathine penicillin G or a similar repository preparation, q 48-72 h. For S/E, C/I, D/F, D/I, W/P, see page 34

- ✓ Amoxicillin or Ampicillin 10 mg/kg q 24 h, IM
- W/ P: Meat 6 days; milk 96 hr.
- Long acting penicillin preparation before dying the cow C/I, S/E,D/I see page 34
- Oxyteracycline 10mg/kg, q 24 h, IV
- W/P: milk 96 hour; meat 14-28 days; For C/I, S/E, D/I, see page 34

Intramammary infusion:

- ✓ Benzathine Cloxacillin, 500mg for 3 days
- S/E: like Benzathine or procaine penicillin (see page 34)
- C/I: hypersensitivity and not recommended for food producing animals
- D/F: Intramammary suspension 500, 625 mg/dose

Prevention

- ✓ Disinfection of the teat before and after milking
- ✓ Dry cow therapy with long acting penicillin preparation should be applied before drying the cow.

3.2.1.4. Lumpy skin disease (LSD)

Lumpy skin disease is an infectious, eruptive, occasionally fatal disease of cattle caused by a virus associated with the nettling poxvirus in the genus Capri poxvirus of the family Poxviridae [54]. Pneumonia is a common sequel in animals with lesions in the mouth and respiratory tract. it is a highly infectious viral disease of cattle and buffalo characterized by boxlike intracutaneous firm nodules, edema of the lips, superficial lymph nodes swelling, and lyphangitis. The disease is widely distributed in Ethiopia imposing severe economic loss due to damage of hide [4].

Clinical symptoms

Painful swelling, fever, nasal discharge, emaciation, and hypersalivation followed by characteristic eruptions on the skin and other parts of the body are characteristics; the nodules are circumscribed, round, slightly raised, firm, and painful and involve the entire Curtis and the mucosa of the gastrointestinal, respiratory and genital tract. The regional lymph nodes are swollen and edema develops in the udder, brisket, and legs. Secondary infection cause extensive suppuration and sloughing lesions [4].

Diagnosis: - The widespread nodular lesion of the skin and mucous membranes and biphasic fever are indicative [4].

Treatment and prevention

- ✓ **Drug treatment:** There is no effective treatment, but secondary bacterial infections are prevented by administration of broad-spectrum antibiotics.
- ✓ **Prevention: -** vaccination with sheep/ goat pox virus or LSD strain.

3.2.1.5. Foot-and-Mouse disease (FMD)

Foot-and-Mouse disease (FMD) is a highly communicable viral infection of cattle, pigs, sheep, goat, buffalo, and artiodactyl wild life species characterized by fever, vesicle in the mouse, on the muzzle, gums, pharynx, teats, and interdigital cleft. It is caused by an Aphthovirus that is transmitted by contact and through milk. Recovered animals remain carriers for up to 2 and a half year [4]. The foot-and-mouth disease is a highly communicable disease affecting cloven-footed animals. The disease spreads by direct contact or indirectly through infected water, manure, hay and pastures [3].

Clinical symptoms: -The clinical sign includes inappetance, fever, shacking or rolling, of the feet, reduced milk yield, high temperature, shivering followed by smacking the lips, and pregnant animals may abort [4].

Diagnosis: - Clinical signs are indicative and confirmed by Foot Mouse Disease (FMD) serology.

Treatment and prevention

- ✓ **Drug treatment: -** No specific treatment, however, supportive treatments against secondary bacterial infection are necessary.
- ✓ **Control and prophylaxis: -** vaccination, test and quarantine infected herds

3.2.1.6. Pneumonia

It is inflammation of the pulmonary parenchyma usually accompanied by inflammation of the bronchioles and often by pleurisy. It is the main respiratory problem of dairy cattle. The occurrence of this problem might be due to stress, workload and movement of animals during drought period that can favors the bacteria to multiply due to the immune status of the animals were suppressed. Environmental risk factors include close confinement and poor ventilation of the house, high temperature that encourage replication of the diseases [55].

Important pathogens associated with pneumonia in cattle are Mannheimia haemolytica serotype A,Haemophilus sommus, P, multocida, Mycobacterium Bevis, Mycoplasma mycoides subspecies mycoides small colony type, Parainfluenza virus 3 bovine respiratory syncytial virus, bovine herpes virus 1 and bovine viral diarrhea [4].

Clinical symptoms

It is manifested clinically by an increase in respiratory rate, cough, salivation, impotence, abnormal breath sounds on auscultation and, in most bacterial pneumonias, by evidence of toxemia. Broncho-pneumonia is usually accompanied by a moist, painful cough; interstitial pneumonia is characterized by frequent, hacking coughs, often in paroxysms. The presence of nasal discharge depends on accompanying inflammation of the upper respiratory tract [4].

Diagnosis

The clinician has to decide whether there is pneumonia and if there is, then determine the nature and cause of the pneumonia [4].

Treatment and prevention

Non-drug treatment

- ✓ The treatment of pneumonia depends on the etiology
- ✓ Nursing: sick animals should be provided with shelter; animals would be given good quality long-stem pastures.

Drug treatment

- ✓ **Procaine pencillin** G, 22000 IU/kg; maintenance q 24 h , PO for 3 to 5 days or Benzathine pencillin G or similar respiratory preparation q 48-72 h. for S/E, C/I, D/I, D/F W/P, see page 34
- ✓ **Ampicillins tetrahydrate** 22 mg/kg IM, SC, q 24 h for 3-6 days. For S/E, C/I, D/I, D/F W/P, see page 34
- ✓ **Oxyteracycline hydrochloride** 11 mg/kg, q 24 h or long acting formulation, 20 mg/kg, IM, q 48 h for 3-5 days. S/E, C/I, D/I, D/F W/P, see page 35
- ✓ **Tylosin** 44 mg/kg, IM, q 24 h for 3-5 days

- S/E: allergic reaction in all species and gastrointestinal disturbance
- C/I: animals with impaired liver function
- D/F: powder 10, 20, and 30%; injection, 50, 200, 150 and 220 mg/ml and tablet 200mg
- D/I: theophylline, warfarin and beta-adrenergic drugs
- W/P: adult, meat 7 days, milk 4 days; calves, meat 14 days

3.2.1.7. Internal Parasite (Gastrointestinal)

Internal parasites are infectious of the gastrointestinal tract with nematodes, cestodes and trematodes. The common stomach worms of cattle are Haemonchus placei, Ostertagia ostertagi and Trichostronglyus axeii. Ostertagia and Trichostronglyus infections are characterized by profuse, watery diarrhea that usually is persist. Signs of anemia, hyperproteinemia and edema, particularly the lower jaw and sometimes along the ventral abdomen manifests these infections together with haemonchus infections [4].

Clinical Symptoms: Ostertagia and Trichostronglyus infections are characterized by profuse, watery diarrhea that usually is persist. Signs of anemia, hyperproteinemia and edema, particularly the lower jaw and sometimes along the ventral abdomen manifests these infections together with haemonchus infections [4].

Diagnosis: animals with poor body condition have anemia and diarrhea is suggestive; confirmed by fecal examination. Clinical signs, grazing history and season may give a presumptive diagnosis of internal parasite infection in cattle. The diagnosis confirmed by finding worm eggs on a fecal exam [4].

Treatment:

Reduce exposure to parasite burdens by moving calves onto paddocks that have not been grazed for a long period. Avoid too much grazing pressure higher pasture covers reduce the number of larvae consumed as most larvae live close to the ground. Calves which are healthy and well fed will be less susceptible to gastrointestinal parasites than weaker calves. Grazing paddocks with adult cattle or sheep after calves will reduce the larvae load on the pastures.

Calves need to be drenched before worm burdens get too high. Can be administered as a pour on, oral or injectable. Oral combination drenches tend to be the most effective drench in young calves. Talk to your vet re avoiding drench resistance - factors include using the correct dose, returning treated stock to contaminated pastures, not treating healthier animals, only drenching when necessary based on faecal eggo counts (FEC) and symptoms.

- ✓ Haemonchus placei: Tetramisole 15 mg/kg
- S/E: frothing, salivation, tremore, transient head shaking, licks of lip, urination, defecation, vomiting, ataxia, collapse and death due to respiratory failure
- C/I: within 14 days of treatment of organophosphorus compound or diethylcarbamate. Don’t exceed dose 4.5 gm per animal
- D/F: Bolus, 150,600,700,1000,1200,1500,and 2000 mg;powder or granule, 10,20, and 30% ; injection, 30 and 100 mg/ml
- ✓ Ostertagia ostertagi:Albendazol 7.5 mg/kg
- ✓ Trichostronglyus axeii: Ivermectin 200 mg/kg

- S/E: Ataxia, depression, tremors, mydriasis, listlessness, msculoskeletal pains, oedema of the faceor extremeties itching and popular rash
- C/I: calves less than 12 weeks of age and lactating animals
- D/F: Bolus, 1.75 gm and 5 gm; Suspension, 800mg/ml
- W/P: meat 28 days and don't use in lactating animals

✓ **Control:** To be effective, a parasite control program must reduce the numbers of worms in all classes of cattle and control the number of worms on the pasture. Control of internal parasites in cattle must kill all stages in the animal and help control the number of larvae and eggs on pastures. Adult cattle usually have more resistance to internal parasites than younger cattle, but deworming older cattle can help reduce pasture contamination.

3.2.1.8. Enteritis /Diarrhea

It is swelling or inflammation of the small intestines and can present as a variety of symptoms in the bovine. Enteritis also called diarrhea is a common in new born calves characterized by progressive dehydration and death, some times in a few as 12 hr. it is caused by multitude of bacterial, viral and non-infectious agents [4].

Clinical Finding:

Subacute form: diarrhea persists for several days and result in malnutrition and emaciation. It is common in diary calves [4].

Acute form: The major signs are diarrhea, dehydration, profound weakness, and death within one to several days of onset. The sign depend on the etiology [4].

Diagnosis: Slight to moderate distension of right abdomen; fluid-rushing and splashing sounds on auscultation and ballottement Diarrhea and dehydration.

Treatment and prevention:

Drug Treatment:

✓ **Sulfadimidine**, initial dose: 140mg/kg, IV; maintenance dose: 70mg/kg, IV q 24 h for5-7 days.

- S/E: crystallization in urinary tract, cutaneous eruption, hypothyroidism and idiosyncratic toxicities.

- C/I: pregnant and lactating animals
- D/I/: thiopentone sodium and warfarin
- D/F: bolus, 5g; injection 330,333 and 160mg/ml; powder, 8,10,16,20,25, and 30%
- WP: meat 21 days; it should not be used in lactating cows

✓ **Procaine pencillin**, G, 22,000IU/kg, IM or SC q 24 h for 3-5 days

✓ **Benzathine pencillin** or similar respiratory preparations, q 28-72 h. S/E, C/I, D/I, D/F W/P, see page 34

3.3. Knowledge Acquired Using Interview

In this study to acquire the needed knowledge, both secondary and primary (documented and undocumented) sources of knowledge are used. Primary knowledge is gathered from veterinaries and pharmacist (five experts), in Basso Animal Health Center by using interviewing and critiquing knowledge elicitation methods. In the same way, secondary sources of knowledge are collected by using document analysis.

Interview (both structured and unstructured) is used to collect tacit knowledge from the domain experts. In addition, critiquing (analyzing) elicitation methods are used to purify the collected knowledge. The acquired knowledge is refined with the consultation of the domain experts. Moreover, secondary sources of knowledge are gathered from the Internet, animal disease diagnosis and treatment guidelines (especially, The Drug Administration and Control Authority of Ethiopia, Standard Veterinary Treatment Guidelines for veterinary clinics), research papers and articles by using document analysis technique.

Purposive sampling method is used to select domain experts for knowledge acquisition. The selection criteria of domain experts for the study are based on the professions/expertise, educational qualification level, year of experience and their immediate position in the cattle disease diagnosis and treatment. Due to this, the researcher selects only three experts.

As the domain experts explained during discussion, there are two types of diagnosis that the veterinarian uses to treat the cattle. These are tentative diagnosis and differential diagnosis.

1. **Tentative diagnosis** is used to treat cattle which are infected with the disease. Some diseases show specific clinical signs when the animal is infected. For example, Lumpy Skin Disease, which is caused by a virus can be diagnosed and treated quickly without comparing symptoms. This is

because, the disease has easily identified symptom. These kinds of diseases can be treated by asking the case history from animal owners.

2. **Differential diagnosis**: This is used to diagnose certain diseases which have two or more similar clinical signs. In this case, in order to identify a disease, the veterinarian should diagnose the difference between the infected animal by using clinical symptoms.

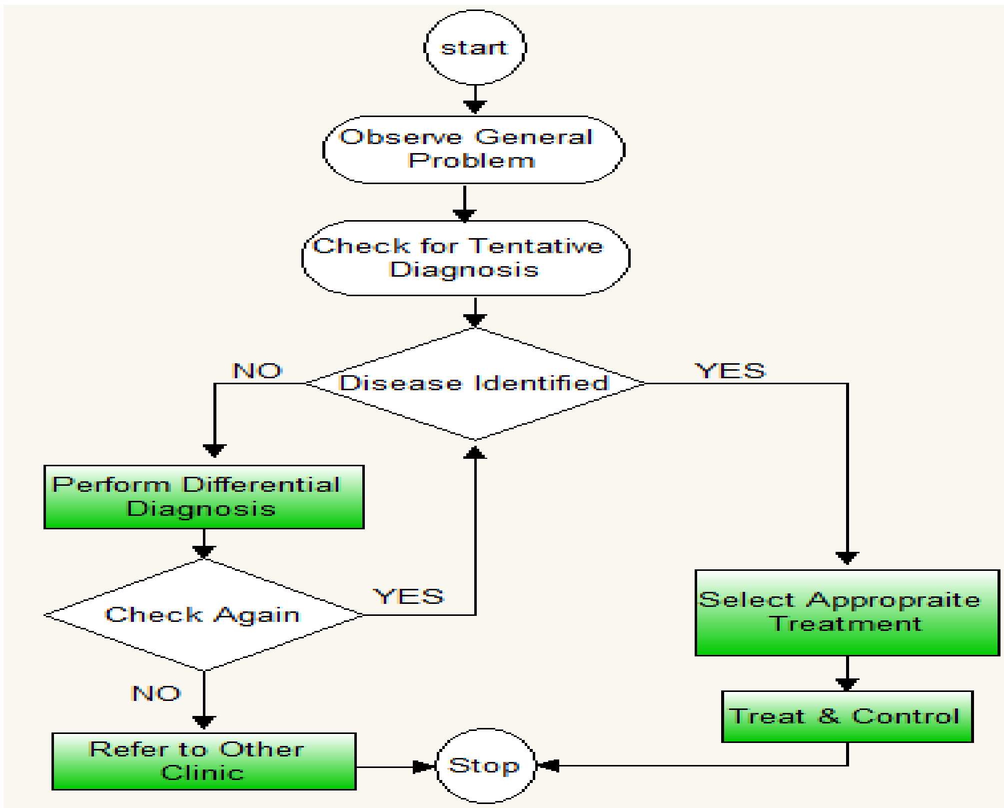


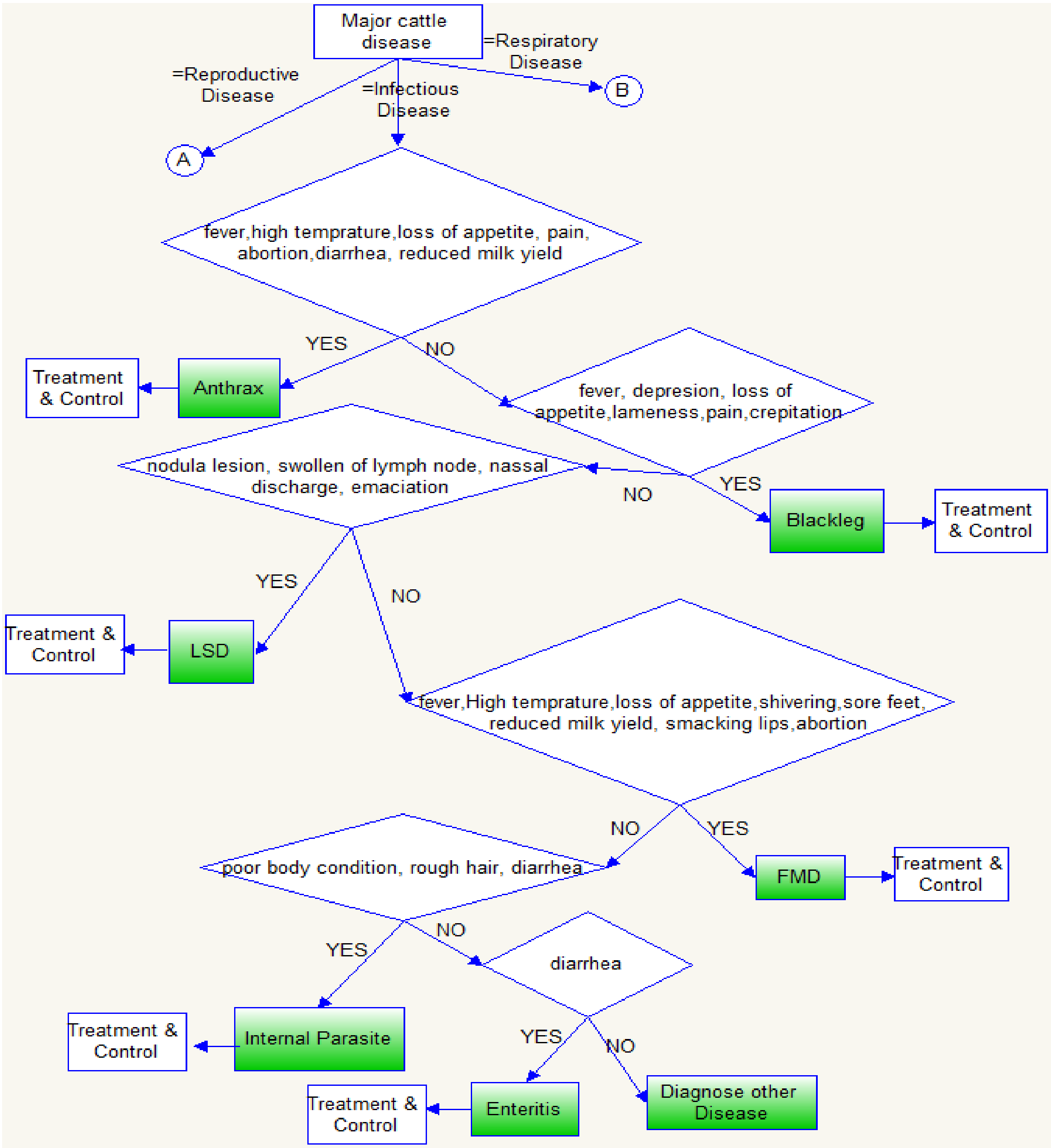
Figure 3.2 Diagnostic Procedure for cattle Disease

The researcher observed how the veterinarians diagnose a particular disease in the infected cattle. From the observation and discussion, the procedure of diagnosis that takes place in the domain is demonstrated in the figure 3.2 above.

3.3.1. Conceptual Modeling for Cattle Disease Diagnosis and Treatment

Cattle disease diagnosis involves many steps beginning with collecting information about the cattle disease owner's case history, symptoms, and performing a clinical examination by veterinaries. Although there are no laboratory tests in the Basso Animal Health Center to diagnose specifically cattle disease, the veterinary use various tests to make sure something else is not causing the symptoms. The veterinary bases his or her diagnosis on the cattle owner case history of symptoms, including any functional problems caused by the symptoms and his or her observation. The veterinary then determines if the case history symptoms and degree of disability point to a diagnosis of a specific disease. Cattle disease diagnosis is one and basic part in the diagnosis of animal medical health conditions which needs careful examination and diagnosis.

In the Clinical Diagnosis, a decision needs to be made whether the cattle disease is suspected or not. The input knowledge role consists of data about the diseased case such as, Fever, Cough, Painful, Diarrhea, Nasal discharge, Abortion, Loss of appetite, Swelling of neck, Lameness, Crepitation sound, Change color of milk, Decrease milk yield, Swelling udder, Shivering, Sore feet, Salivation, Rolling, Smacking lips, Depression, Nodular lesion, Abnormal breathing, Rough hair, Emaciation and Poor body condition. The following figure 3.3 shows Decision tree for cattle disease from discussion of domain expert:


Major cattle disease
=Reproductive Disease
=Infectious Disease
=Respiratory Disease
A
B
fever,high temprature,loss of appetite, pain, abortion,diarrhea, reduced milk yield
YES
NO
Treatment & Control
Anthrax
fever, depression, loss of appetite,lameness,pain,crepitation
nodula lesion, swollen of lymph node, nassal discharge, emaciation
NO
YES
Blackleg
Treatment & Control
YES
NO
Treatment & Control
LSD
fever,High temprature,loss of appetite,shivering,sore feet, reduced milk yield, smacking lips,abortion
NO
YES
poor body condition, rough hair, diarrhea
FMD
Treatment & Control
YES
NO
diarrhea
Treatment & Control
Internal Parasite
YES
NO
Treatment & Control
Enteritis
Diagnose other Disease

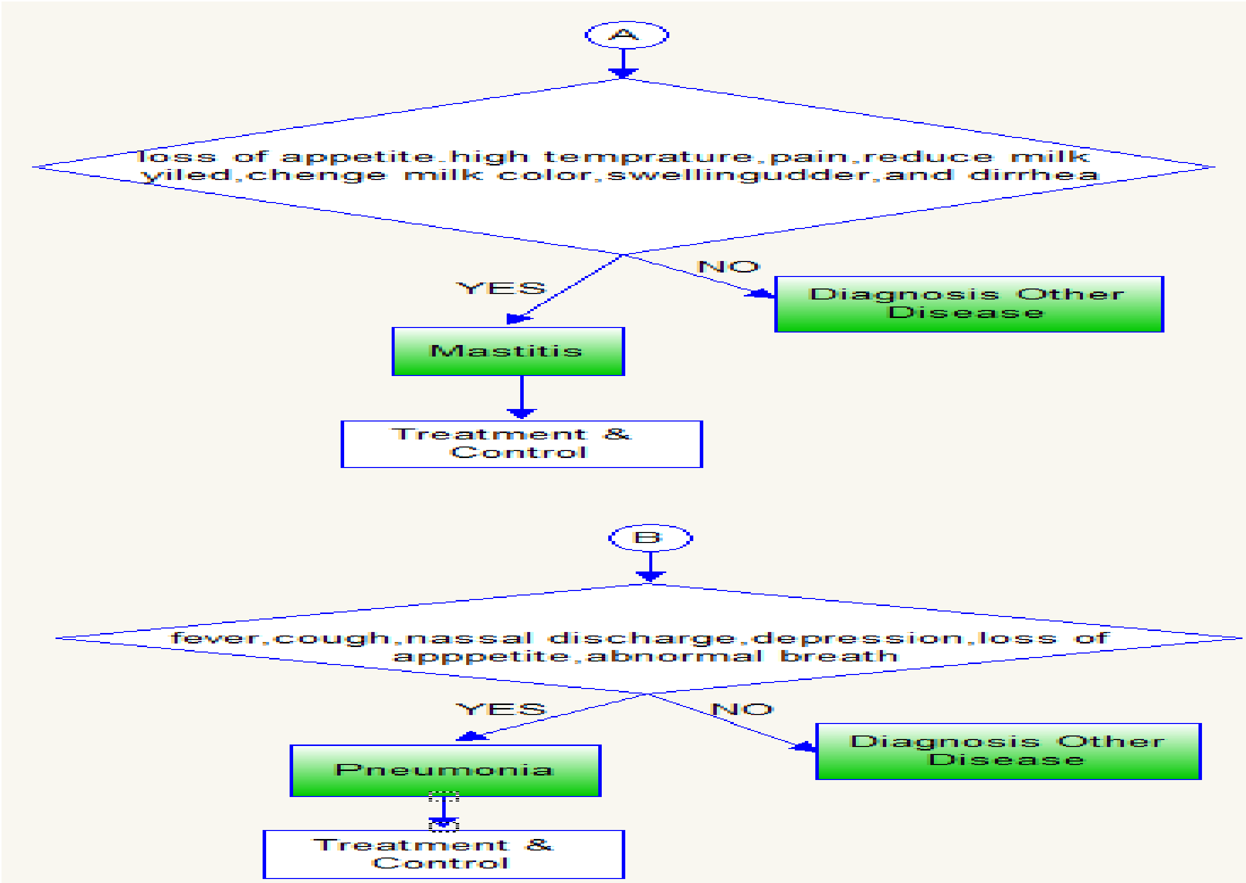


Figure 3.3 Decision tree for cattle disease from discussion of domain expert

3.4. Knowledge Acquired from Data Mining

Nowadays, data stored in medical databases are growing in an increasingly rapid way. Due to this tendency data mining application in healthcare today is excessive, because healthcare organizations today are capable of generating and collecting a large amounts of data. This increase in volume of data requires an automatic way for these data to be extracted when needed. With the use of data mining techniques it is possible to extract interesting and useful knowledge and this knowledge can be used by physicians to determine diagnoses, prognoses and treatments for patients in healthcare organizations which improve work efficiency and enhance the quality of the decision making process [56]. The dataset for this study have been collected from Basso Animal Health Center, which is found in Debre Birhan, North Shoa Ethiopia. In chapter four discuss in detail about the data mining.

CHAPTER FOUR

DATA MINING METHODOLOGY

4.1. Introduction

This section discusses about methodology i.e. data mining techniques used through the research process which includes general approach of the research design, the data collection, analysis and interpretation of artifacts. Having the description of the type of the study follows a hybrid model knowledge discovery approach which starts from understanding of the problem to use discovered knowledge in six steps iteration.

4.2. Research Design with Respect to Hybrid Model

In order to apply the Data mining technology, one must follow the standard process steps, for instance hybrid model can be taken. Below here each step was summarized for this study. Figure 4.1 illustrates system framework as follows:

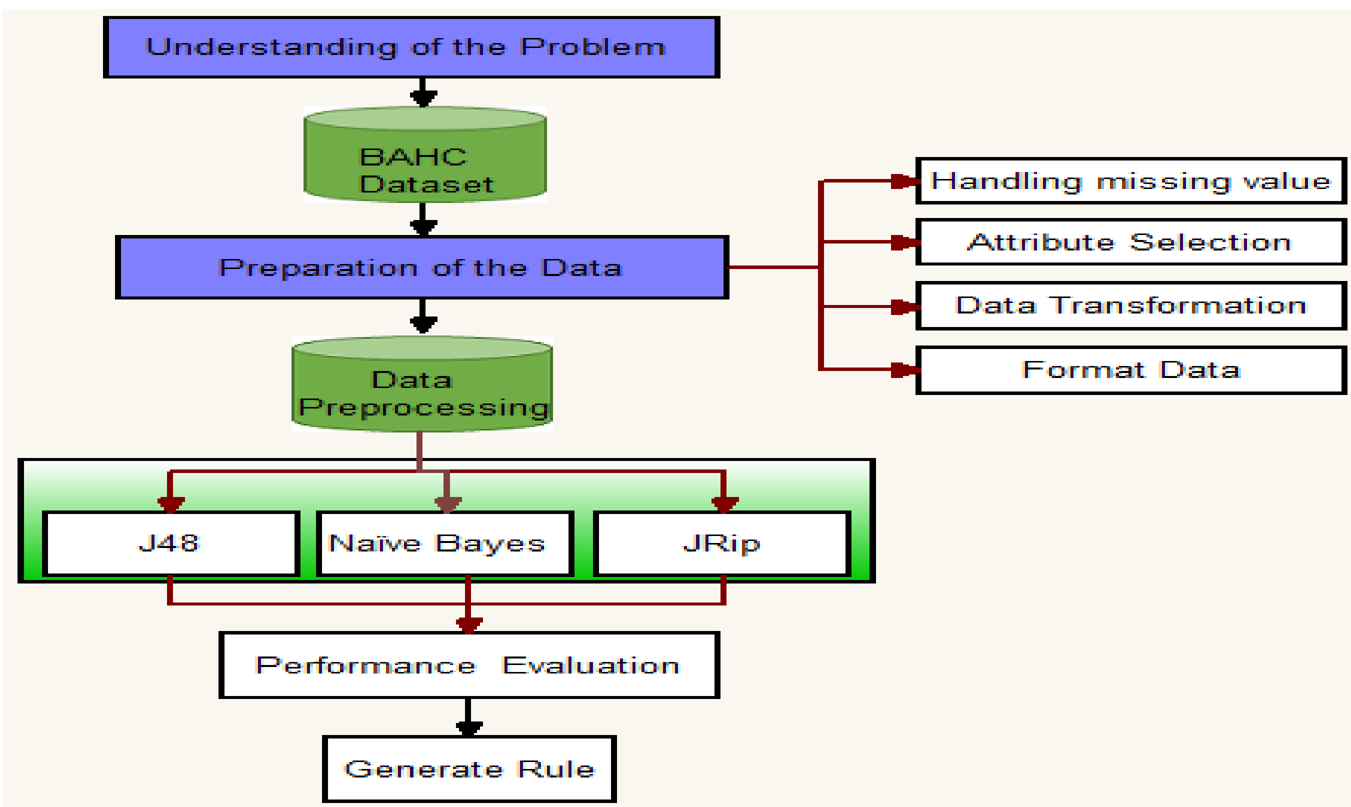


Figure 4.1 Data Mining System framework

4.2.1. Understanding of the Problem Domain

The researcher understands the problem domain by conducting some structured and unstructured sample interview questions for purposely selected domain experts. This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, like an animal health officer, veterinaries, laboratory technician, pharmacist and learning about the current solution to the problem. Here is a significant number of cattle are infected by disease before diagnosis and treatment of the disease. Finally, the project goal is translated into data mining goals, and the initial selection of data mining tools to be used later in the process is performed [17]. In this case since the animal health center is demanding the answer of how to increase the performance of diagnosis and treatment of diseased cattle's? A better classification technique at the beginning is very much essential in order to learn from the stored data. An overall problem domain is understood via through and repeated discussing with stakeholders identified above and reviewing an abstract journal which is prepared by the animal health Bureau of the region.

4.2.2. Understanding of the Data

This step includes collecting sample data and deciding which data, including format and size can be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, etc. finally, the step includes verification of the usefulness of the data with respect to the data mining goals [17]. There is no database in Debre Birhan Animal Health Center (DBAHC). However, it uses a manual document to record animal disease, symptoms, and treatment of the disease every day throughout the year. And this data is collected from Debre Birhan Basso Animal Health Center (DBBAHC). Initially the data were manual table format and then to write each record data into Microsoft Excel 2016 for further preparation and getting ready for input to WEKA 3.8 data mining tool. The initial data are collected from BAHC for a time period of two years. Collecting total number of records from the area is 2175 with 33 attributes. The total size of the initial data in Excel format is 1.96 MB.

During the data understanding Phase the application process for cattle disease diagnosis and treatment at the Basso animal health center is studied, including the interviews from the veterinaries' and guidelines, in order to identify the types of data collected from the diseased animal registered and stored in the Basso animal health center databases in manual format.

4.2.3. Preparation of the Data

This step covers all activities needed to construct the final dataset, which constitutes the data that can be fed into data mining tool(s) in the next step. To be useful for data mining purpose, the database need to undergo preprocessing, in the form of data Cleaning, data reduction, data transformation and Waikato Environment for Knowledge Analysis (WEKA) understandable format.

4.2.3.1. Handling Missing Values

BAHC dataset has instances with total records of 2175 and 33 attributes including class label, among these records 123 (7.7%) are missing values in only temperature attribute. We use Matlab as a tool to handle missing values by using mean average techniques to fill the missing values.

This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs. Let us consider that the value xij of the k-th class, Ck, is missing then it will be replaced by:

$$x^{\wedge ij} = \sum_{i:xij \in CK} xij/nk \ldots \ldots \ldots \ldots \ldots \ldots (2.16)$$

Where nk represents the number of non-missing values in the j-th feature of the k-th class.

Table 4.1 Sample data after handling missing values

Tempratu	Fever	Coughing	Painful	Diarrhoea	Nasal Dis	Abortion	Loss of Ap	Swell of n	Lameness	Crepitati	Change c	Decrease	Swelling	Shivering	Sore feet	Salivation	Rolling	Smaking	Depressi	Nodular L	UbNorma	Rough ha	Emaciatio	poor Body	Age	Class
39.5	Yes	No	No	No	No	Yes	No	No	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	Adult	FMD
39	No	No	No	No	No	No	Yes	No	Yes	Yes	No	No	No	Yes	No	No	Yes	No	Yes	No	No	No	No	No	Adult	Blackleg
42.3	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	No	No	Yes	No	No	No	No	No	No	Yes	No	No	No	No	No	Adult	Anthrax
39.5	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	Young	LSD
42.3	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	No	Yes	No	No	No	Adult	Pneumonia
38.5	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Adult	Enteritis
38.5	Yes	No	No	No	No	No	No	No	Yes	Yes	No	No	No	Yes	No	No	Yes	No	No	No	No	No	No	No	Adult	Blackleg
38.9	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Adult	Enteritis
39.1	No	Yes	No	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	Adult	Blackleg
39.5	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	Adult	LSD
38.5	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Adult	Internal Parasite
38.5	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Calf	Enteritis
39.5	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	Adult	LSD
42.3	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	No	Yes	No	No	No	Adult	Pneumonia
39.5	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	No	No	Yes	No	No	No	No	No	No	Yes	No	No	No	No	No	Adult	Anthrax
41.2	Yes	Yes	Yes	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	Young	Pneumonia
37.5	No	No	No	No	No	No	No	No	Yes	Yes	No	No	No	Yes	No	No	Yes	No	No	No	No	No	No	No	Adult	Blackleg
42.3	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	No	Yes	No	No	No	Young	Pneumonia

4.2.3.2. Data Transformation

Data transformation, the data are transformed or consolidated into forms appropriate for mining [24]. In this study, the researcher performed generalization of the records by using categories for only temperature attribute which is grouped in low, normal and high. In the initial dataset, the body temperature is measured using degree centigrade (^{0}C) and there are more than 39 the high value, the range between 38 and 39 normal and less than 38 low. So, the researcher categorized them into three groups as indicated in table 4.2.

Table 4.2 Data Transformation by the cattle temperature range *[4]*

Value	Temperature in degree centigrade (^{0}C)
Low	<38
Normal	38-39
High	>39

4.2.3.3. Format Data

Formatting transformations refers to primarily syntactic modifications made to the data that do not change its meaning, but might brokered by the modeling tool. After all preprocessing the data is covered into Dot Comma Separated Value (**. CSV**) format and for the WEKA it's again converted into dot attribute Relation File Format (**. ARFF**) file format. Figure 4.2 illustrates sample WEKA ARFF files show as follows:

```
@relation 'wekafil-weka.filters.supervised.instance.SMOTE-C0-K5-P10.0-S1-weka.filters.supervised.instance.SMOTE-C0-K5-P10.0-S1-weka.filters.supervised.

@attribute 'Swell_of_ neck' {No,Yes}
@attribute Swelling_Udder {No,Yes}
@attribute Decrease_Milk_yield {No,Yes}
@attribute Temprature {High,Normal,Low}
@attribute Smaking_Lips {Yes,No}
@attribute Change_color_of_Milk {No,Yes}
@attribute Abnormal_Breathing {No,Yes}
@attribute poor_Body_condition {No,Yes}
@attribute Coughing {No,Yes}
@attribute Nodular_Lesion {No,Yes}
@attribute Lameness {Yes,No}
@attribute Rough_hair {No,Yes}
@attribute Crepitation_Sound {No,Yes}
@attribute Sore_feet {Yes,No}
@attribute Depression {No,Yes}
@attribute Diarrhoea {No,Yes}
@attribute Fever {Yes,No}
@attribute Painful {No,Yes}
@attribute Nasal_Discharge {No,Yes}
@attribute Loss_of_Apptetite {No,Yes}
@attribute Class {FMD,Blackleg,Enteritis,LSD,Pneumonia,'Internal Parasite',Anthrax,Mastitis}
```

Figure 4.2 ARFF file from WEKA

4.2.3.4. Class Imbalance

The class imbalance problem is prevalent in many applications, including: fraud/intrusion detection, risk management, text classification, and medical diagnosis/monitoring, etc. It typically occurs when, in a classification problem, there are many more instances of some classes than others. In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. Particularly, they tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class [57].

A number of solutions to the class-imbalance problem were proposed both at the data and algorithmic levels. At the data level, these solutions include many different forms of re-sampling such as over-sampling and under-sampling. Synthetic Minority Over-sampling Technique (SMOTE) is an over-sampling approach which generates synthetic examples in a less application specific manner. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors [57]. Figure 4.3 illustrates imbalance dataset before SMOTE shows as follows:

Selected attribute

Name: Class | Type: Nominal

Missing: 0 (0%) | Distinct: 8 | Unique: 0 (0%)

No.	Label	Count
1	FMD	232
2	Blackleg	441
3	Enteritis	342
4	LSD	336
5	Pneumonia	406
6	Internal Parasite	195
7	Anthrax	97
8	Mastitis	126

Class: Class (Nom) | Visualize All

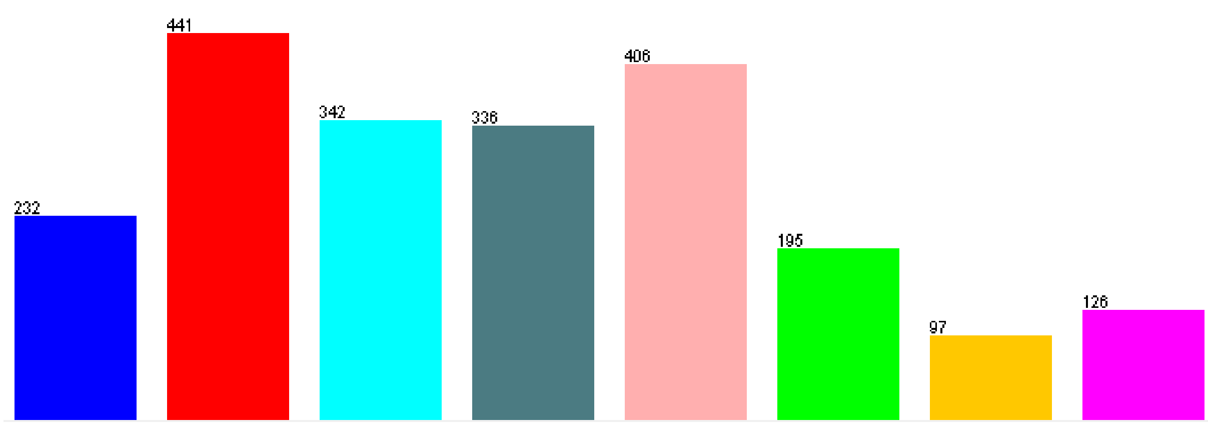


Figure 4.3 Imbalance dataset before SMOTE

As we can observe in figure 4.4 below, the classes are balance and to avoid imbalance, the researcher SMOTE the dataset and discretize it before conducting the experiments and as a result the dataset increases from 2175 records to 3627 records. Add 1452 instances in the real data instances to increase for the purpose of balancing each class label.

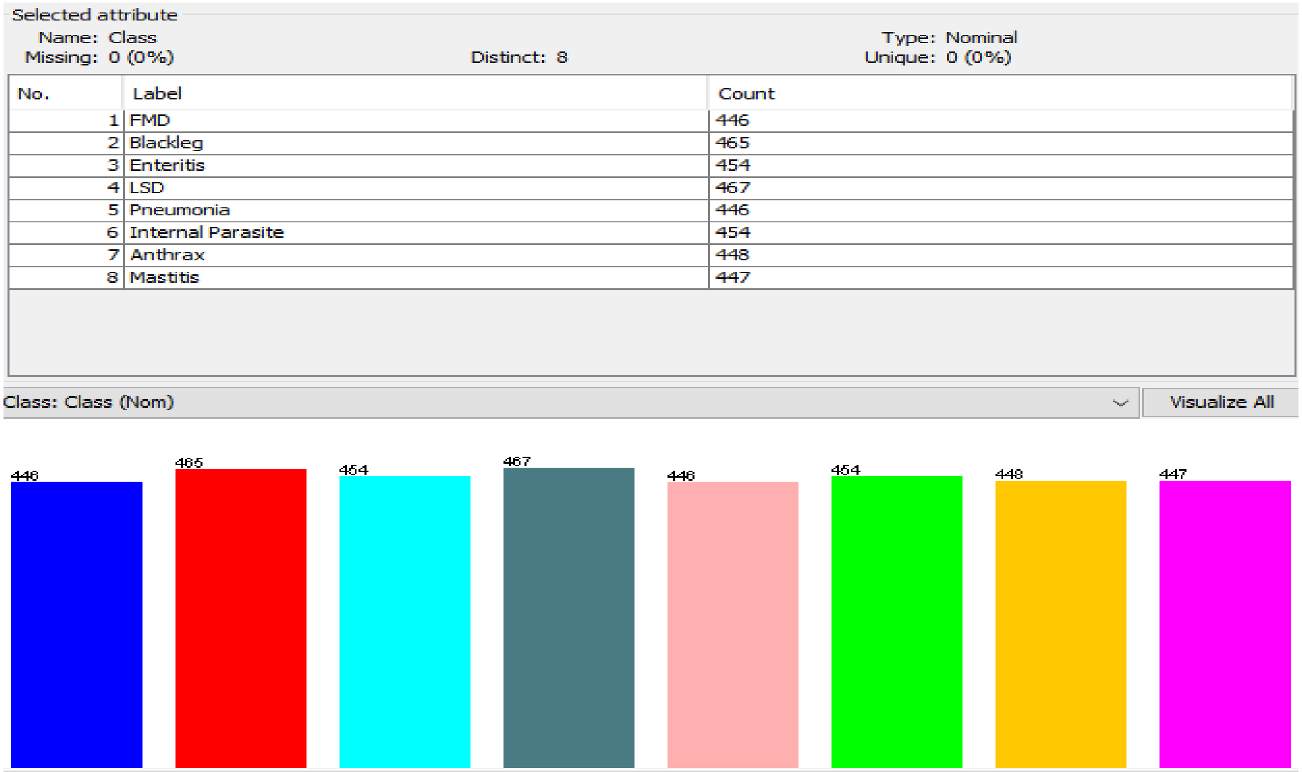


Figure 4.4 Balanced Dataset after SMOTE and Discretize

4.2.3.5. Attribute Selection

Attribute selection searches through all possible combinations of attributes in the data and fines which a subset of attribute works best for prediction. Attribute selection also known as feature selection which is select relevant attributes and remove redundant and/ or irrelevant attributes. In this research, roll number, date and kebele, removed because all are not important to determine the cattle disease. Species, sex and breed also removed because all are only identifiers of animals. We want to eliminate irrelevant attributes and evaluate attributes individually, so that we use the ranking method with GainRatioAttributeEval were followed for selecting best attribute selection for data mining that come up 20 out of 33 attributes. Table 4.3 shows unnecessary attribute for diagnosis and treatment for cattled disease as follows:

Table 4.3 Removing unnecessary attributes of the table

Removed attributes	
Attribute name	Remark
Roll No, Date, Kebele, Spp, Sex, Breed	Not important for the study

Attribute Evaluator (supervised, Class (nominal): 21 Class):

Gain Ratio feature evaluator

Ranked attributes:

0.861 8 Swell_of_ neck
0.853 13 Swelling_Udder
0.848 12 Decrease_Milk_yield
0.846 1 Temperature
0.842 15 Smaking_Lips
0.815 11 Change_color_of_Milk
0.811 18 Abnormal_Breathing
0.808 20 poor_Body_condition
0.801 3 Coughing
0.783 17 Nodular_Lesion
0.782 9 Lameness
0.779 19 Rough hair
0.771 10 Crepitation_Sound
0.77 14 Sore_Feet
0.748 16 Depression
0.742 5 diarrhea
0.612 2 Fever
0.609 4 Painful
0.598 6 Nassal_discharge
0.443 7 Loss_of_Apptetite

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

The experiment shows that the attribute a Swell of neck, Swelling Udder, Decrease Milk yield, Temperature and Smacking Lips are the five most impact of the output and Painful, nasal discharge and Loss of Appetite have the three smallest output impact.

CHAPTER FIVE

KNOWLEDGE EXPERIMENT USING DATA MINING

5.1. Introduction

In this chapter, discuss the experimentation process by relating the steps followed, the choice made, the task accomplished, the result obtained, evaluation of the model results, and present it in a way that the organization can easily understand and use it. And also to discuss means to question the research findings, and to consider different interpretations. Experiments have been carried to develop classifier model and to extract relevant actionable classification rules. In this study different experiments were conducted using various data mining methods to derive knowledge from preprocessed data for diagnosis of cattle disease from the preprocessed data. According to the methodology of this study after preparation of the data, the next task is the mining process. As it has been stated in the previous sections, a total of 3627 data records were preprocessed to perform the experiment.

5.2. Selecting Modeling Technique

Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of these research three classification techniques, namely the Naïve Bayes classification, decision tree classification and rule based classification has been selected for model building. The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, Naïve Bayes, J48 and JRip algorithms are used for experimentation of this study. The reason why the researcher selected the above algorithms is, because the algorithms are easy to understand and interpret the results of the model, and also have advantages such as high tolerance to noise, and the ability to classify unseen patterns.

5.3. Experimental Setup

In any data mining research before developing a model, we should generate a mechanism to test the model performance. For instance, in the supervised data mining task, such as classification, it is common to use classification accuracy measure, True Positive rate (TP), precision, recall and F-measure of the experts are used as to measure the performance of the developed data mining model.

5.3.1. Methods of Training and Testing Option

For the experimental setup, the cattle disease datasets were converted to ARFF (Attribute Relation File Format), because it is a suitable input file format for the WEKA system. In order to perform the experiment the researcher used two methods to classify the dataset, the first method is k-fold (10-folds) cross validation and, the second method is percentage split.

Therefore the dataset is randomly partitioned equally into ten parts. Consequently, 90% of the dataset is used for training and 10% for testing and the dataset are partitioned into percentages splits option (70%: 30%) that means 70% of the dataset is used for training and the remaining 30% for testing purpose. As the researcher explained on the experimental setup to build the model for each algorithm perform three experiments based on the two methods, namely the experiment performed based on K-fold (10 – fold) cross validation method called experiment I and the second experiment performed based on the percentage split method called experiment II..

5.3.2. Developing Classifier Model Using Naïve Bayes

Experiment I:

This experiment conducts under 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as Naïve Bayes and Correctly Classified Instances are 3572 which means 98.48 % and Incorrectly Classified Instances are 55 which means 2.15% from Total Number of 3627 of Instances and taking 0.02 seconds to build the model.

Experiment II:

Percentage split test option to train and test the classification model. Out of the 3627 total records 2539 (70%) of the instances were used as training dataset and the remaining, 1088 (30%) of the instances were used as a testing dataset. The Naïve Bayes learning algorithm scored an accuracy out of total 1088 number of testing instances 1070 (98.35 %) of them are classified correctly and the remaining 18 (1.65%) testing instances are misclassified or incorrectly classified.

To conclude, the above two experiments namely experiment I and II performed in order to build the classification model using Naïve Bayes classification algorithm by applying k-fold cross validation and percentage split method in respectively on the experiments.

Experiment I and Experiment II shows that the Classification accuracy of the models based on the above two methods respectively. The first experiment was performed based on 10-fold cross validation method and classifies with 98.48%, accuracy rate, and the second experiment performed based on 70%:30% percentage split classifies with 98.35% accuracy rate.

So to sum up, when we compared the two experiments the first experiment performed based on K-fold cross validation dataset has a better accuracy performance than the second experiment performed by percentage split testing dataset. Each equations refer from page 28:

Table 5.1 Detailed Accuracy by Class for Naïve Bayes classification algorithm:

Detailed Accuracy by Class					
	TP Rate	Precision	Recall	F-Measure	Class
	97.3%	99.5%	97.3%	98.4%	FMD
	99.6%	97.7%	99.6%	98.6%	Blackleg
	99.1%	94.7%	99.1%	96.9%	Enteritis
	100%	97.9%	100%	98.9%	LSD
	96.9%	99.8%	96.9%	98.3%	Pneumonia
	98.0%	99.1%	98.0%	98.6%	Internal Parasite
	100%	99.6%	100%	99.8%	Anthrax
	96.9%	100%	96.9%	98.4%	Mastitis
Weighted Ave	98.5%	98.5%	98.5%	98.5%	

5.3.3. Developing Classifier Model Using J48 Decision Tree

Experiment I:

This experiment conducts under 10-fold cross-validation test option with default parameters of Weka and the algorithm generates a model as a decision tree with 35 number of Leaves and 68 size of the tree and correctly classified Instances are 3575 which means 98.65% and Incorrectly Classified Instances are 49 which means 1.35% from Total Number of Instances of 3627 and taking 0.08 seconds to build the model..

Experiment II:

This experiment conducted in percentage split test option to train and test the classification model. Out of the 3627 total records 2539(70%) of the instances were used as training dataset and the remaining, 1088(30%) of the instances were used as a testing dataset. The J48 learning algorithm scored an accuracy out of 1088 total number of testing instances 1069 (98.25%) of them are classified correctly and the remaining 19(1.75%) testing instances are incorrectly classified. The algorithm generates a model as a discussion tree with 35 number of leaves and 68 size of the tree and taking 0.06 seconds to build the model.

To conclude, the above two experiments namely experiment I and II performed in order to build the classification model using J48 classification Algorithm by applying k-fold cross validation and percentage split methods in respectively on the experiments. The first experiment was performed based on 10-fold cross validation method and classifies with **98.65 %** accuracy rate, and the second experiment performed based on 70%:30% percentage split classifies with **98.25%** accuracy rate.

So, when we compared the two experiments performed by different methods, the first experiment performed based on K-fold cross validation has a better accuracy performance than the second experiment performed by percentage split. Each equations refer from page 28:

Table 5.2 Detailed Accuracy by Class for J48 classification algorithm:

Detailed Accuracy by Class					
	TP Rate	Precision	Recall	F-Measure	Class
	98.7%	98.2%	98.7%	98.4%	FMD
	98.3%	98.9%	98.3%	98.6%	Blackleg
	98.9%	97.2%	98.9%	98%	Enteritis
	100%	97.7%	100%	98.8%	LSD
	96.9%	99.5%	96.9%	98.2%	Pneumonia
	97.6%	99.1%	97.6%	98.3%	Internal Parasite
	99.6%	99.6%	99.6%	99.6%	Anthrax
	99.3%	99.1%	99.3%	99.2%	Mastitis
Weighted Ave	98.6%	98.6%	98.6%	98.6%	

5.3.4. Developing Classifier Model Using JRip Rule Based

Experiment I:

In this experiment JRip rule induction algorithm is employed. Therefore, to generate IF-THEN rules from the experimental BAHC dataset JRip algorithm with its default values of the parameter and 10-fold cross-validation test mode is employed. JRip correctly classified 3579 which means

98.68% instances and incorrectly classified 48 which means 1.32% instances. The algorithm generates 21 rules taking 0.27 seconds to build the model.

Experiment II

For this experiment the K-fold cross validation method is changed into percentage split test option to train and test the classification model. Out of the 3627 total records 2539 (70%) of the instances were used as training dataset and the remaining, 1088 (30%) of the instances were used as a testing dataset. The JRip learning algorithm scored an accuracy out of 1088 total number of testing instances 1069 (98.25%) of them are classified correctly and the remaining 19 (1.75%) testing instances are misclassified or incorrectly classified. The algorithm generates 20 rules and taking 0.30 seconds to build the model.

To conclude, the above two experiments namely experiment I and II performed in order to build the classification model using JRip classification Algorithm by applying k-fold cross validation and percentage split methods in respectively on the experiments. The first experiment was performed based on 10-fold cross validation method and classifies with **98.68** accuracy rate, and the second experiment performed based on 70%:30% percentage split classifies with 98.25 accuracy rate. So, when we compared the two experiments performed by the two methods, the first experiment performed based on K-fold cross validation has a better accuracy performance than the second experiment performed by percentage split. Each equations refer from page 28:

Table 5.3 Detailed Accuracy by Class for JRip classification algorithm:

Detailed Accuracy by Class					
	TP Rate	Precision	Recall	F-Measure	Class
	99.8%	98.5%	99.8%	99.1%	FMD
	97.4%	99.6%	97.4%	98.5%	Blackleg
	98.7%	97.2%	98.7%	97.9%	Enteritis
	99.6%	96.5%	99.6%	98.0%	LSD
	96.4%	99.5%	96.4%	97.9%	Pneumonia
	98.0%	98.9%	98.0%	98.5%	Internal Parasite
	99.8%	99.6%	99.8%	99.7%	Anthrax
	99.8%	100%	99.8%	99.9%	Mastitis
Weighted Ave	98.7%	98.7%	98.7%	98.7%	

5.3.5. Performance Comparison of Naïve Bayes, J48 and JRip Models

Selecting a better classification technique for building a model, which performs best in handling the classification, is one of the aims of this research. For this reason, the three selected classification model with respective best performance accuracy is listed in below table 5.4:

Table 5.4 Comparison of the results of the models

Algorithms used	Correctly classified instances		Incorrectly classified instances		Time taken (second)
10 Fold Cross Validation Test Option					
	No.	Accuracy	No.	Accuracy	
Naïve Bayes	3572	98.48 %	55	1.52%	0.02
J48	**3578**	**98.65%**	**49**	**1.35%**	**0.08**
JRip	3579	98.68%	48	1.32%	0.27
Percentage Split Test Option					
Naïve Bayes	1070	98.35 %	18	1.65 %	0.02
J48	1069	98.25 %	19	1.75 %	0.06
JRip	1069	98.25 %	19	1.75%	0.03

In these experiments three algorithms are used, namely Naïve Bayes, J48 decision tree and JRip rule based. From each methods totally six models are developed based on the two methods. As shown on the above comparison table (Table 5.4), all the results were almost all closely equal, but the difference lies in the execution period or the time taken to build the model.

The graphical representation of the algorithms with respect to classes for TP rate is indicated in Figure 5.1:

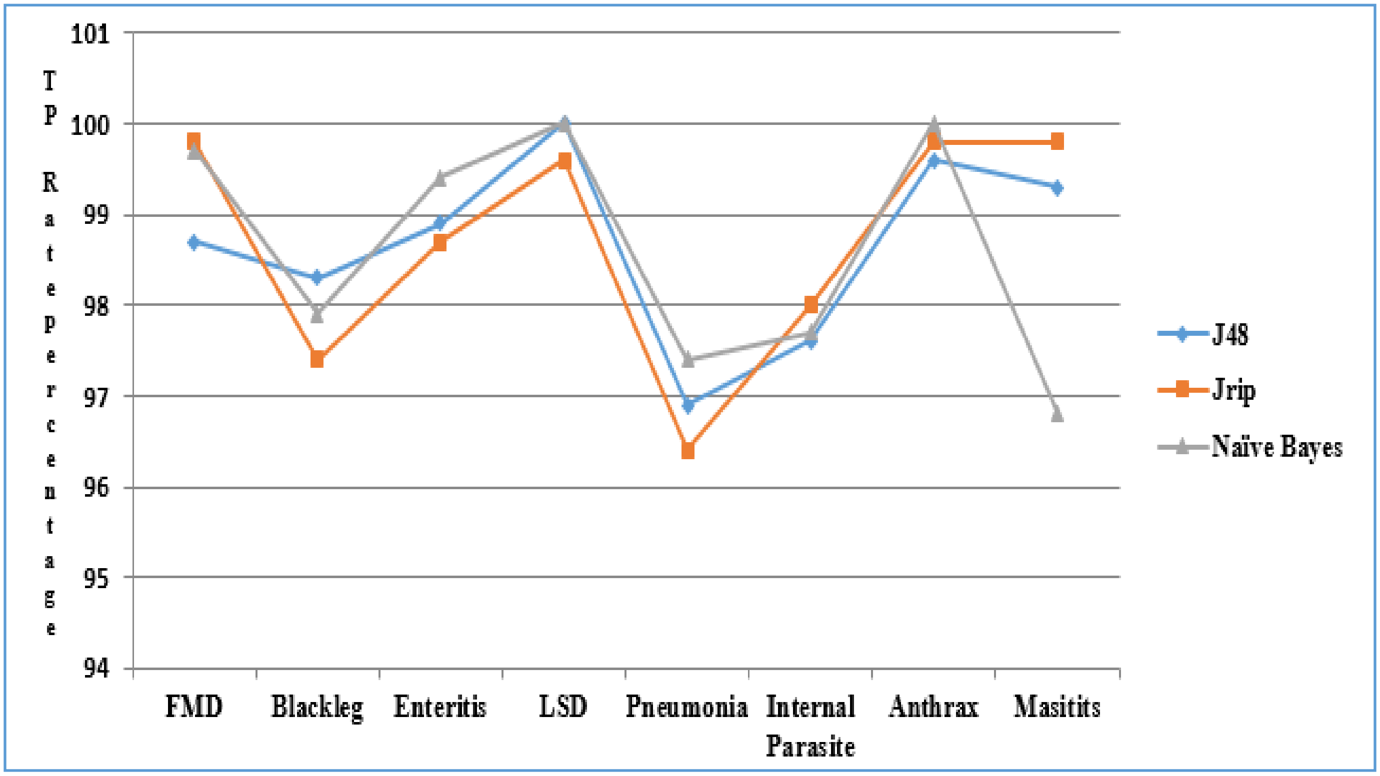


Figure 5.1 TP rate of classifiers

5.4. Classifying Model Performance Evaluation Metrics

Evaluation of classifier data mining algorithms can be compared according to a number of measures. In comparing the performance of different classifier data mining algorithms to determine its classifications, some quantities that interpret the goodness of fit of a model, and error measurements must be considered.

The final comparative analysis of the models shows on the above table 5.4 JRip rule classifier algorithm with 10-fold cross validation test option are performed best classification accuracy of 98.68%. Whereas J48 and Naïve Bayes classifier algorithms with 10-fold cross validation test option performed classification accuracy of 98.65% and 98.48% of the result respectively.

In other side, the final comparative analysis of the models shows on the above table 5.4 the JRip classifier algorithm and J48 decision tree have equal Percentage split test option in the testing dataset performed classification accuracy of 98.25%, whereas Naïve Bayes algorithm with percentage split test option performed best classification accuracy of 98.35% result.

Finally, the final comparative analysis of the models shows on the above table 5.4 the Naïve Bayes classifier algorithm with Percentage split test option in the training dataset performed best classification accuracy of 98.62%, whereas J48 and JRip classifier algorithms with percentage split test option training dataset performed classification accuracy of 97.44 % and 97.65% result respectively.

So, to sum up the model which was developed by the JRip rule classifier algorithm with 10-fold cross validation test option method was selected as the best classification model based on the accuracy evaluation methods used in this study.

Hence, to evaluate the performance of the classifiers employed in this study True Positive rate, Precision, Recall and F-measure are used.

5.5. Rule Extraction from JRip Classification Algorithm

Having generated rules using JRip classifier, the next task is building or constructing the knowledge base. For this study, we devised an automatic construction of knowledge bases aligned with the data mining task. The overall task of the application is to extract rules from BAHC dataset using JRip classifier and mapping JRip rules to Prolog rules.

Table 5.6 Rules extracted by JRip Classification algorithm

S.No	Rule
1	(Coughing = Yes) => Class=Pneumonia (352.0/0.0)
2	(Abnormal_Breathing = Yes) and (Depression = No) => Class=Pneumonia (74.0/0.0)
3	(Temprature = Normal) and (Fever = Yes) and (Decrease_Milk_yield = No) and (Nodular_Lesion = No) and (Painful = No) and (Loss_of_Apptetite = Yes) and (Lameness = No) and (Depression = No) => Class=Pneumonia (7.0/1.0)
4	(Smaking_Lips = Yes) => Class=FMD (385.0/2.0)
5	(Sore_feet = Yes) => Class=FMD (34.0/0.0)
6	(Temprature = High) and (Painful = No) and (Loss_of_Apptetite = No) and (Swelling_Udder = No) and (Crepitation_Sound = No) and (Fever = Yes) and (Depression = No) => Class=FMD (28.0/3.0)
7	(Lameness = Yes) and (Crepitation_Sound = No) and (Painful = No) => Class=FMD (4.0/1.0)
8	(Swelling_Udder = Yes) and (Nodular_Lesion = No) => Class=Mastitis (400.0/0.0)
9	(Change_color_of_Milk = Yes) => Class=Mastitis (33.0/0.0)
10	(Temprature = High) and (Fever = No) and (Crepitation_Sound = No) and (Painful = No) => Class=Mastitis (14.0/0.0)

11	(Decrease_Milk_yield = Yes) => Class=Anthrax (409.0/0.0)
12	(Swell_of_ neck = Yes) => Class=Anthrax (34.0/0.0)
13	(Depression = Yes) and (Nasal_Discharge = Yes) and (Painful = Yes) => Class=Anthrax (5.0/0.0)
14	(Diarrhoea = Yes) => Class=Enteritis (401.0/0.0)
15	(Fever = No) and (poor_Body_condition = No) and (Nodular_Lesion = No) and (Temprature = Normal) and (Rough_hair = No) and (Nasal_Discharge = No) and (Loss_of_Apptetite = No) => Class=Enteritis (56.0/9.0)
16	(poor_Body_condition = Yes) => Class=Internal Parasite (430.0/5.0)
17	(poor_Body_condition = Yes) => Class=Internal Parasite (430.0/5.0)
18	(Rough_hair = Yes) and (Temprature = Normal) => Class=Internal Parasite (20.0/0.0)
19	(Temprature = High) => Class=Blackleg (444.0/0.0)
20	(Loss_of_Apptetite = Yes) and (Fever = Yes) and (Nasal_Discharge = No) => Class=Blackleg (14.0/0.0)
21	=> Class=LSD (483.0/16.0)

CHAPTER SIX

INTEGRATION OF DATA MINING WITH KNOWLEDGE BASED SYSTEM

6.1. Introduction

Agent based techniques are now used to reason and coordinate knowledge discovery tasks across distributed data repositories; neural networks are now used to optimize data mining parameters; spectral clustering is now used in Cased Based Reasoning (CBR) for medical discovery, and incremental learning is know the means to 'idiot-proof' business intelligence systems. Both disciplines have carried this potential to be used in a closed loop fashion, where the reasoning of AI helps to 'soften' the problems brought about force analytics of data mining. And data mining turn, is the key to producing the relevant models and patterns that AI algorithms require [58]. In this study for the integration of data mining with knowledge based system weak module for swi-prolog called SWIWEKA1.0 is used and all the steps taken to make the integration are discussed in this chapter.

6.2. Contents of SWIWEKA

To do this, Java programming was used to integrate WEKA result with the Knowledge Based System automatically. And also 'swiweka' is used as an interface that allows the use of the WEKA API for classification; weka.jar, Weka _src.jar is used to construct a model when called from interface through swiweka package, a JPL library to connect the Java layer with the Prolog layer. The interface contains the following Prolog programs:

JPL_weka.pl: - this is a bunch of intermediate tools used to make easier the handling of WEKA using JPL.

Dataset.pl: - this is the main part of the interface. It mainly wraps the weka.core.Package of weka to deal with datasets (Instance, attribute and instance object)

Fastvector.pl: - just few predicates to handle the weka.fastVector class

arffFiles.pl:- Input/output operations over ARFF file formats (the basic format supported by WEKA)

resources.pl: - analyzer of the WEKA APP to deal with its element

classifiers.pl:- common predicates to deal with classifier objects

Runclassification.pl: common calls to run classifier

Weka. pl: - this is a module that provides quite a rough set of predicates to interact with the WEKA APP and deal with ARFF files. This module is loaded by the classification front end, which provides a framework to use it properly. This module defines the general and more important predicates.

6.3. Installing SWIWEKA

SWIWEKA is Linux based software and installing the Linux operating system is necessary for the best performance of the package. So, in this study, the researcher used Ubuntu 14.04.5. After installing the operating system the prolog files listed in section 6.2 are downloaded from the google code swiweka folder and saved in one folder. Then downloaded the Linux package from the WEKA website and copy weka.jar and weka_src.jar into the swiweka folder and set the classpath at the bash by opening .bashrc file for the jar files as indicated in listing 6.1:

```
Sudogedit.bashrc
```

Listing 6.1. Linux command for opening the .bashrc file

6.4. Integrated Knowledge Based

While integrating the knowledge base with the data mining results the prolog code that performs the integration operation is discussed below in sequence.

6.4.1. Calling Java from Prolog

Prolog (programming in logic) is one of the most widely used programing languages in artificial intelligence (AI) researchers. As opposed to imperative language such as C or Java (the latter of which also happens to be object-oriented) it is a declarative programing language. That means, when implementing the solution to a problem, instead of specifying how to achieve a certain goal in a certain situation, we specify what the situation (rules and facts) and the goal (query) are and let the prolog interpreter derive the solution for us. Prolog is very useful in some problem areas, such as AI, neural language processing, database, etc., but pretty useless in others, such as graphics or numerical algorithms [59]. The expressiveness of prolog is due to three major features: rule based programming, built in pattern matching and backtracking execution. The rule based programming allows the program code to be written in a form which is more declarative than procedural. This is made possible by the built-in pattern matching and backtracking which automatically provide for the flow of control in the program. Together, these features make it possible to elegantly implement many types of knowledge based system [60]. So, by using prolog the researcher tries to call Java class. For this task SWI-prolog platform has JPL, which is a bidirectional Java/prolog package. The knowledge base uses SWI-prolog JPL library to connect the Java layer with the prolog layer. The library module JPL provides functionality for controlling the loading and unloading the JVM (Java Virtual Machine), method call functionality (jpl_call/4), and predicates for managing object references.

```
Jpl_call (+X, +method, +args, -R)
```

Where X can be a type, class object or classname (for static method of the denoted class, or for static or instance methods of java.lang.class). Also, it can be a class instance or array (for static or instance methods). The Method can be an atomic method name (if this name is ambiguous, as a result of method overloading, then it will be resolved by considering the type of Args, as far as they can be inferred). Args must be a proper list (possibly empty) of ground arguments and finally, an attempt will be made to unify R with the returned result [61]. In this study the jpl_call is used to call writeToSample method in writeToARFF, this method is used for writing the values of each attribute at the end of ARFF file, which are entered by the user at the prolog prompt.

```
Jpl_call('writetoarff.WriteToArff',writeToSample,[Temperature,Smacking_lips,
sorre_Feet,Fever,Lameness,Depression,Painful,Crepetation_sound,Loss_of_Appetite,Diarrhea,N
odular_lesion,Nasal_discharge,Coughing,Abnormal_breath,Swell_of_udder,Decrease_milk_yiel
d,Change_color_of_milk,Swell_of_neck,Rough_hair,Poor_body_condition],_)
```

Listing 6.2 jpl_call calling the java class weka JRip coding from prolog

6.4.2. The Java Class

In this study the Java class is used to change the values entered by the user into the setting and write them into dataset ARFF which is used for storing instance.

```
Public static void writeToSample (string Temperature, string Smacking_lips, string Sore_Feet,
string Fever, string Lameness, string Depression, string Painful, string Crepetation_sound, string
Loss_of_Appetite, string Diarrhea, string Nodular_lesion, string Nasal_discharge, string
Coughing, string Abnormal_breath, string Swell_of_udder, string Decrease_milk_yield, string
Change_color_of_milk, string Swell_of_neck, string Rough_hair, string Poor_body_condition)
```

Listing 6.3. The writeToSample method in writeToARFF class

In the writeToSample method the pass values are changed into a string using if---else conditional statements and then this statement is written at the end of the dataset ARFF as shown in line 2 of listing 6.4.

```
1. bufferedWriter writeEntry=new BufferedWriter(new FileWriter("Tadeset.arff", true));
2. writeEntry.write(strtemp+ "," +strsmaking+ "," + strsore + "," + strfever +"," +strlame+
   ","+strdepression+ "," + strpain + "," + strsound +"," +strappetite+ "," +strdiarrhea+ "," + strlesion
   + "," + strdischarge +"," +strcough+ "," +strbreath+ "," + strudder + "," + strmilkyield +","
   +strmilkcolor+ "," +strneck+ "," + strhair + "," + strcondition +"," +"FMD"+ "\n");
3. writeEntry.close();
```

Listing 6.4. The writing capability of writeToSample method

6.4.3. Reading the ARFF Files

In this study two arff files are used, one for storing attribute values entered by the user and the other for the original dataset. So both of them need to read for the proper implementation of knowledge base system. As shown in listing 6.5 wkpl_read_arff creates a dataset by reading the ARFF file. While creating the dataset the wkpl_read_arff predicate calls the wkpl_open_arff predicate and jpl_call.

```
1. Wkpl_read_arff(Arffpath, Instance):-
2.         Wkpl_open_arff(Arffpath,ArffLoader),
3.         Jpl_call (ArffLoader, getDataSet,[], Instances).
```

Listing 6.5. Wkpl_read_arff from arffFiles.pl

The wkpl_open_arff predicates call six predicates and one of them is jpl_call. This predicate takes ARFF path which is the path indicating the ARFF file used for opening and the ARFF loader object from weka.core.converters.ArffLoader class and this performed through wkpl_getObjectpredicate (indicated in listing 6.5) which used to retrieve an object from the WEKA API built by the default constructor and indicated in line 2 of listing 6.6 line 3 and 4 of listing 6.6 are convenience predicates to call the constructor of the WEKA object to pass arguments types and values. At line 7 of listing 6.6 the jpl_call pass the ARFF file as an argument for the setSource method found in weka.core.converters. ArffLoader class. Then jpl_call found at line 3 of listing 6.5 is used to get the dataset and pass it through Instances.

```
Wkpl_open_arff(ArffPath, ArffLoader):-
     Wkpl_getObject('weka.core.converters.ArffLoader',ArffLoader),
     Wkpl_new_argsType_array(1,Args),
     Wkpl_add_type_to_args('java.lang.String',Args, 0),
     Jpl_datums_to_array([ArffPath], Values),
    Wkpl_getObject('javaio.File',[Args],[Values],ArffFile),
    Jpl_call (ArffLoader,setSource,[ArffFile],_).
```

Listing 6.6. Wkpl_open_arff from arffFiles.pl

```
Wkpl_getObject(Name,Object):-
     Wkpl_weka_loader(Loader),
     Jpl_call(Loader, loadClass,[Name],Class),
     Jpl_call (Class, newInstance, [], Object).
```

Listing 6.7. Wkpl_get_Object predicate from jpl_weka.pl

As indicating in listing 6.8 below the wkpl_read_arff predicate is used to read both the sample and main ARFF files and put them in Tadeset.

```
Wkpl_read_arff('Tadeset', Mainarff)
```

Listing 6.8 Usage of wkpl_read_arff in the knowledge base system

6.4.4. Setting Class Index and Classifier

For the purpose classification the class index of the class attribute need to be set and SWIWEKA set the class index with wkpl_set_classIndex predicate which have one jpl_call used for passing the instance object the index argument to the setClass Index methods as indicated in listing 6.9 and its usage in the knowledge base system is indicated in listing 6.10.

```
Wkpl_set_classIndex(Instance,Index):-
      Jpl_call (Instances, setclass_index,[Index],_).
```

Listing 6.9: The wkpl_set_classIndex predicate found in dataset.pl

```
Wkpl_set_classIndex (Mainarff, 20),
```

Listing 6.10 Usage of wkpl_set_classIndex in the knowledge based system

In SWIWEKA the classifier is set using wkpl_classifiers predicate which is used to set the classifier as indicated in listing 6.11. this predicate contain one jpl_datums_to_array predicate and one jpl_call and passes two values; the first one is Option which are WEKA classification options like –t indicating training file, the second one is classifier a return value from the weka.classifiers. Classifier class holding the selected classifier.

```
Wkpl_classifier(Options, Classifier):-
        Jpl_datums_to_array(options, Args),
        Jpl_call (‘weka.classifiers.Classifier’,[Args],Classifier).
```

Listing 6.11 The wkpl_classifiers found in classifier.pl

As described as the evaluation section, the model selected for the next steps of the research is JRip rule algorithm With 21 rules. So, the usage of wkpl_classifiers using JRip rule is indicated in listing 6.12.

```
Wkpl_classifier(‘weka.classifiers.rule.JRip’,Classifiers)
```

Listing 6.12 Usage of wkpl_classifiers in the knowledge based system

6.4.5. Building the Classifier

After being certain about preferring JRip, and setting it accordingly the next step is building the classifier using wkpl_build_classifier. This classifier loads a dataset in the classifier passing a weka.Instances object that is read from an ARFF file as indicated in listing 6.13.

```
Wkpl_classifier(Dataset, Classifier):-
        Jpl_call (Classifier, buildClassifier, [Dataset], _).
```

Listing 6.13 The wkpl_build_Classifiers predicate found in classifier.pl

In the knowledge based system wkpl_build_Classifiers is used to as shown in listing 6.12. The Mainarff dataset containing instances created using wkpl_read_arff predicate as shown in line 1of listing 6.8 and Mainarff files a dataset created from the original ARFF file so this file is used for the classifier. Classifier is created in listing 6.12 using JRip rule algorithm.

```
Wkpl_build_classifier (Mainarff, Classifier)
```

Listing 6.14 Usage of wkpl_build_classifiers in knowledge based system

6.4.6. Getting and Classifying the Instance

After building the classifier the next step is to use the classifier for classification. However, before performing classification the instance to be classified should be prepared. As explained in section 6.4.2 the values entered by the knowledge based system user are Witten into dataset ARFF file. When it is added to ARFF file and read using wkpl_read_arff the values are changed into instances. So, the next task getting this instance the first thing is to know the number of instances in the dataset ARFF because the Java class writes the value at the end of the ARFF file. The number of instances can be obtained using wkpl_numberInstances predicate as shown in listing 6.15.

```
Wkpl_number_Instances(Instances, Number):-
        Jpl_call (instances, numInstances,[],Number).
```

Listing 6.15 The wkpl_numberInstances predicate found in dataset.pl

Since the wkpl_numberInstances is used to predict the class value of the given instance of the user and this instance is found in the dataset ARFF file the predicate uses Dataset as shown in listing 6.16.

```
Wkpl_numberInstances (Dataset, Number)
```

Listing 6.16 Usage of wkpl_numberInstances in the knowledge based system

After getting the number of instances this number need to indicate the last instance of the file so the decrement predicate is used for decrementing the number of instances by one and store this value into index variable. This decrement variable is not found in SWIWEKA package and added by the researcher for the use this study.

By using index variable and wkpl_get_instance predicate the researcher retrieve the instance to be predicated from the dataset as shown in listing 6.18. This predicate accepts the dataset and the index, then returns the instance value in the form of WEKA understandable format.

```
Wkpl_get_instance(Instances, Index, I):-
        Jpl_call (Instances, instance, [Index], I).
```

Listing 6.17: The wkpl_get_instance predicate found in dataset.pl

```
Wkpl_get_instance (Repoarff, Index, Instoclassify)
```

Listing 6.18: Usage of wkpl_get_instance in the knowledge based system

The last step that comes is classifying the instance based on the classifier built. And this is done using wkpl_classify_instance which classifies an instance ones the classifier has been built.

```
Wkpl_classify_instance (Instance, Classifier, Result):-
        Jpl_call (Classifier, classify_instance, [Instance], Result).
```

Listing 6.19: The wkpl_classify_instance predicate found in classifier.pl

```
Wkpl_classify_instance(Instoclassify,Classifier, Prediction)
```

Listing 6.20: Usage of wkpl_classify_instance in the knowledge based system

CHAPTER SEVEN

IMPLEMENTATION AND DISCUSSION OF RESULTS

7.1. Introduction

Findings of data mining are the base for the development of knowledge base system through the Integrator application. Facts we get from Data mining can be represented as a rule in the knowledge base. After knowledge acquisition is done using a JRip rule induction algorithm, which performs on data collected from Basso Animal Health Center the facts extracted is represented in the knowledge base system. The main challenge here is how to use his knowledge extracted from data mining for knowledge based system.

7.2. Architecture of the System

Classification algorithms are widely used in various medical applications. Data classification is a two phase process in which first step is the training phase where the classifier algorithm builds a classifier with the training set of tuples and the second phase is classification phase where the model is used for classification and its performance is analyzed with the testing set of tuples [62].

This system was designed with the progression of conceptual design that refined the systems' architecture. Of course, the conceptual design was essential to stabilize the architecture of data mining result with knowledge based system for cattle diseases diagnosis and treatment. Throughout design iterations, the design of data mining result with knowledge based system for cattle diseases, diagnosis and treatment was expended into system architecture to ensure that it supported the cattle disease and treatments. Figure 7.1 shows the overall system design and framework of integrating data mining induced hidden knowledge about cattle disease diagnosis and treatment based on Basso animal health center dataset with knowledge based system.

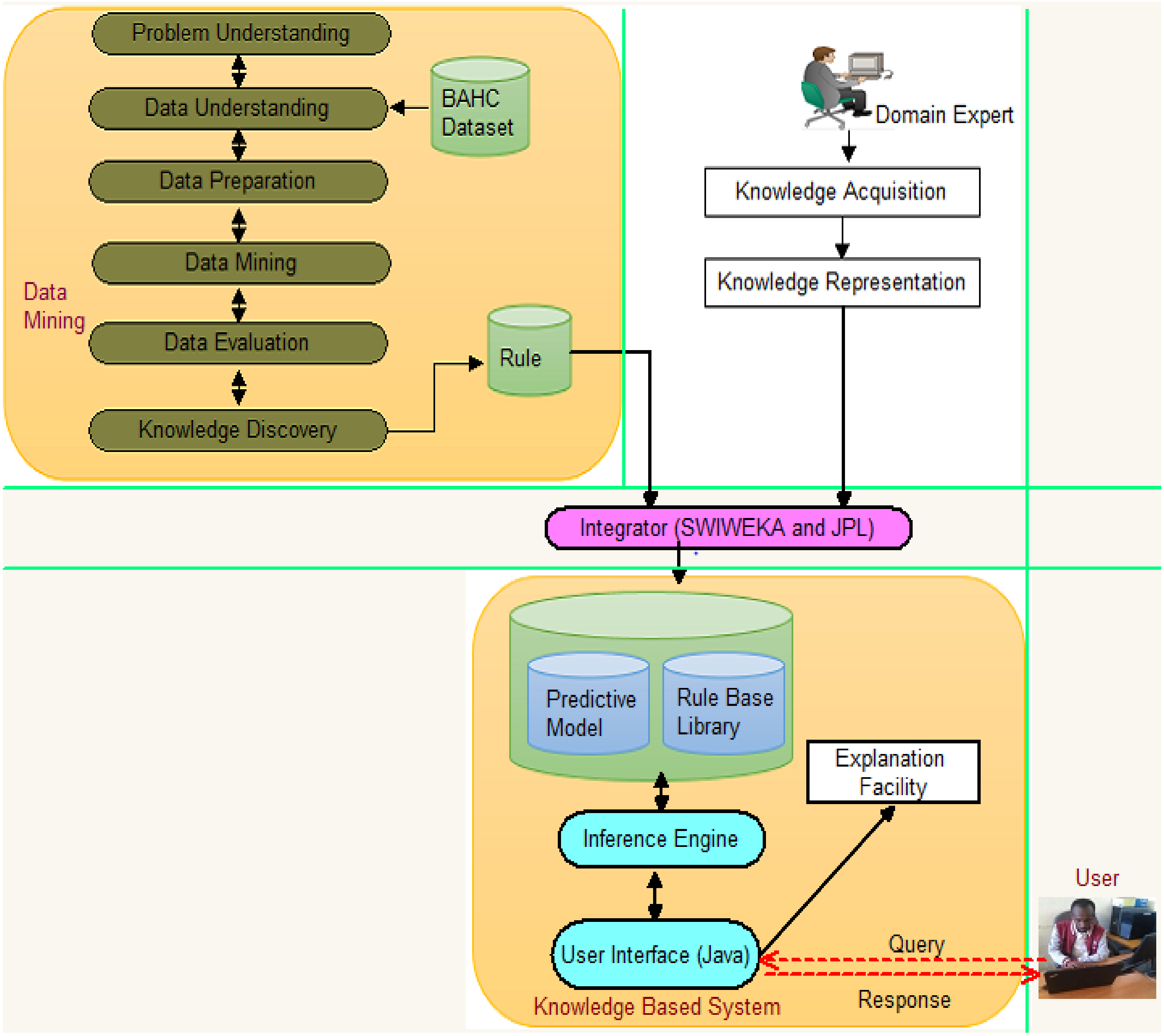


Figure 7.1. General framework Integrating of Data Mining model with KBS [63].

As presented in figure 7.1, the architecture proposed in this research comprises two major components, namely, disease diagnosis and disease treatment. For constructing disease diagnosis model, data mining technology is used to acquire hidden knowledge about the behavior of different disease from the given dataset. For treatment purposes tacit and explicit knowledge is collected from domain experts and manuals and represented.

Data Mining: - Data preprocessing subtask is applied for effective data selection to enhance the performance of the cattle disease diagnosis and treatment system. It contains the feature extraction component to extract the required features from BAHC dataset for the process of disease diagnosis and treatment. During the preprocessing stage of instances are classified as normal and diseased cattle, and also diseased cattle are further classified into eight sections represented into Blackleg, Anthrax, Foot Mouse disease, Pneumonia, Lumpy skin disease, Mastitis, enteritis, and internal parasite.

Integrator: - The JRip model is integrated with knowledge based system automatically for designing intelligent diagnosis and treatment of cattle disease system.

Rule Library: - Is a container of rules about cattle disease which are generated by a JRip rule induction algorithm after mapped by the Integrator to Prolog understandable format.

Knowledge based system: - Let us discuss components of the knowledge Based System for Diagnosis and Treatment of cattle disease as follows: Common component of the system consists of the following subsystems: rule library, Inference engine, knowledge base and user interface.

Knowledge Base: - The researcher stores all knowledge that is collected from domain experts, document analysis and data mining in the knowledge based as set of rules using a rule-based knowledge representation method.

Inference Engine: - An inference engine is the brain of the Knowledge Based System, which directs the system how it can derive a conclusion by looking for possible solutions from the knowledge base and recommend the best possible solution. Since the objective of the proposed Knowledge Based System is for diagnosis and treatment of cattle disease and the Prolog's built-in inference mechanism is backward chaining, the researcher prefers to use a backward inference mechanism which is a goal derived that tries to prove or disprove the goal.

For instance, if the user wants to write a prescription for the diseased cattle after he/she has diagnosis the diseased cattle, he/she should answer the questions that are asked by the system. To do so, the system needs facts and rules about the drugs which are appropriate for the specific diseased cattle.

User Interface: - KBS cattle disease diagnosis and treatment used a simple user interface to display the information. When KBS cattle disease diagnosis and treatment identified a possible type of cattle disease, the system would present treatments and preventions of the disease on the screen and the users can ask detail or more explanation if they did not satisfied with the provided suggestions in such away the users could observe it straight away. For example, when the type of disease was identified, the typical disease descriptions, treatments and prevention would appear on the computer screen and asked whether they need more explanation or not. The user could view different type of disease and treatments for the cattle disease. The first page of the user interface welcomes user's home page as shown in the figure below figure 7.2:

Figure 7.2 Welcoming windows of KBS cattle Disease Diagnosis user interface above

After the prototype displays the home page a user can interact directly with the system by clicking what you want. The following figure shows the dialogue windows between the user and the system to identify the infected cattle.

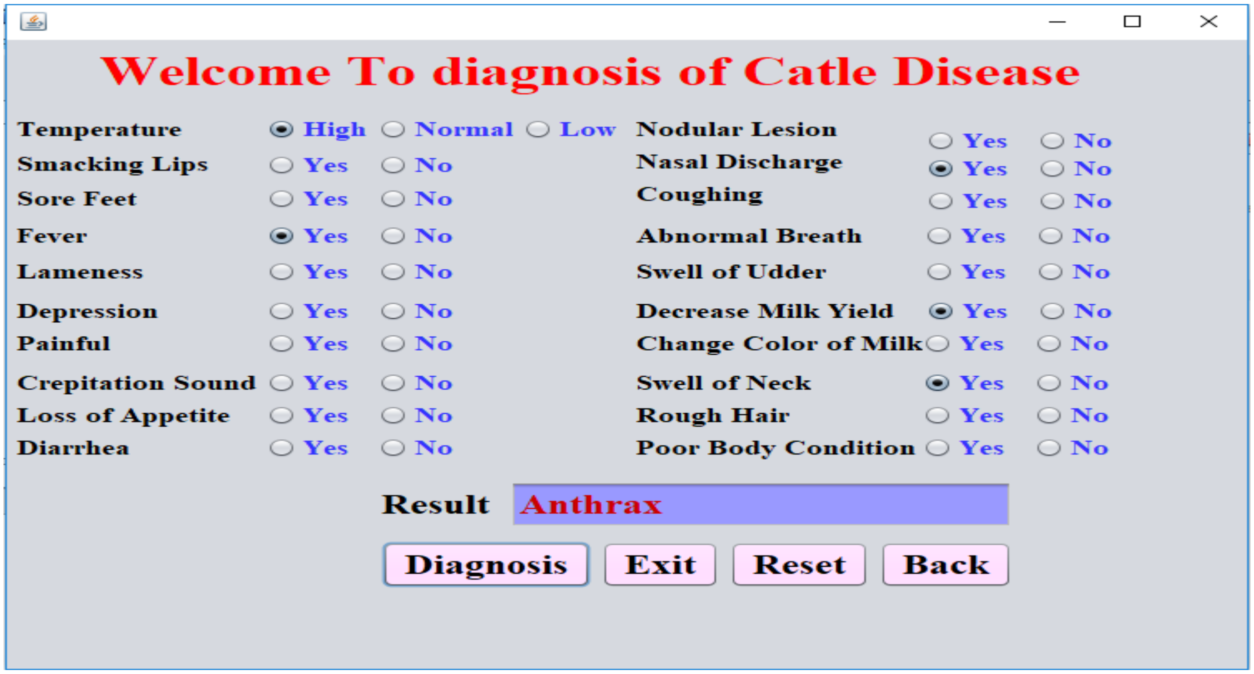


Figure 7.3 Sample windows between the user and the system to identify cattle disease

After the disease is identified and described well its recommended treatments are provided as it is presented in the following sample window.

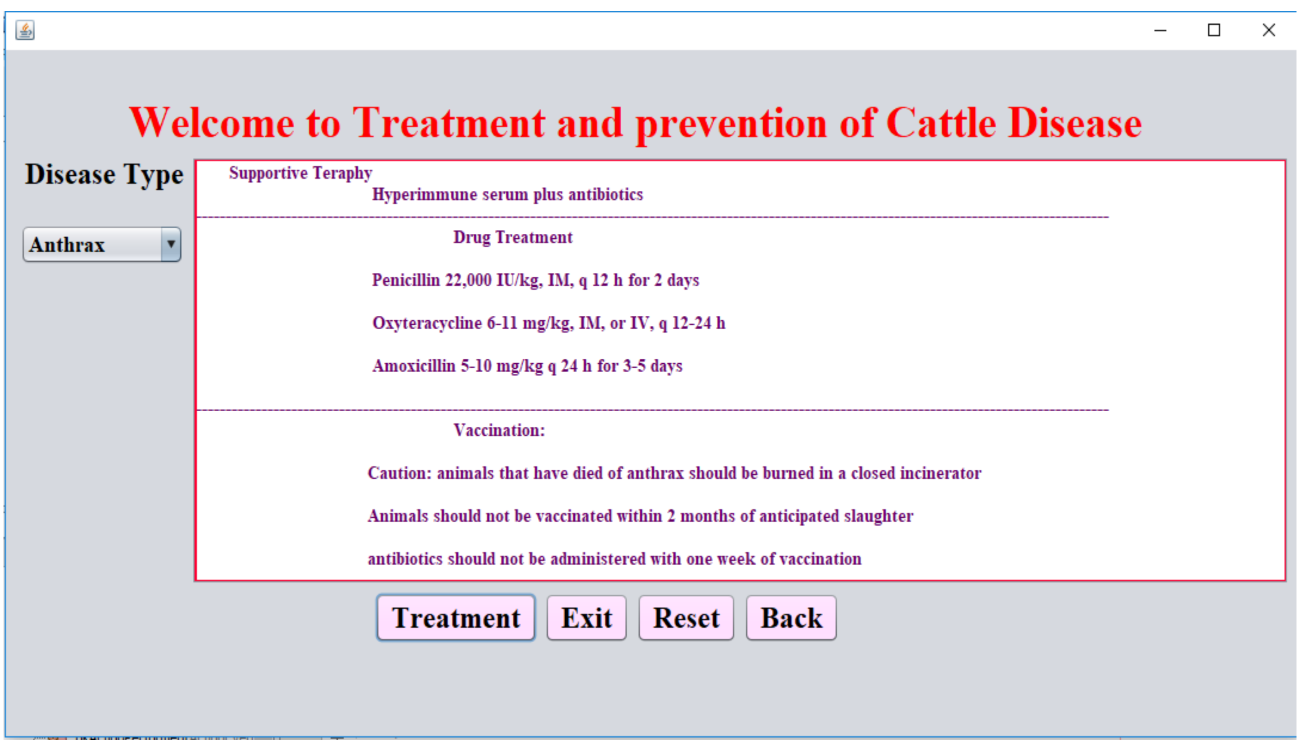


Figure 7.4 Sample window of the system's treat and prevent for the identified disease

7.3. Combination of Data Mining and Knowledge Based System

Accordingly, for diagnosis and treatment of cattle disease the model constructed by JRip rule based is used for cattle disease diagnosis. Java programming was used to integrate diagnosis and treatment of the cattle disease model created using a JRip rule with the Knowledge Based System automatically. For this study, we used Weka, Jpl and swi-prolog tools to construct a predictive model.

7.3.1. Structure of JRip Rule and PROLOG Rule

JRip algorithm generates rules in the form of (condition)… (Conclusion) format. The algorithm generates 21 rules from the BAHC dataset. The format of the some of the rules is indicated in the table 7.1. The condition part contains attribute, a comparison operator and value. Two or more conditions are joined by = and'. After the conditions the '=>' meaning implies follows. The concluding part of the rule has the format class='disease', for example class=Pneumonia, class= FMD.

Table 7.1 Sample JRip rules for Mastitis and FMD.

S.No	Rule	
	Condition	Conclusion
1	(Swelling_Udder = Yes) and (Temprature=High) =>	Class=Mastitis (400.0/0.0)
2	(Temperature = High) and (Painful = No) and (Loss_of_Apptetite = No) and (Swelling_Udder = No) and (Crepitation_Sound = No) and (Fever = Yes) and (Depression = No) =>	Class=FMD (28.0/3.0)

As shown in table 7.1 the rules are in IF…. THEN format. For example:

(Swelling_Udder = Yes) and (Temprature=High) => Class=Mastitis (400.0/0.0)

This rule can be read us:

IF (Swelling_Udder = Yes) and (Temprature=High = No) =>**THEN** class Class=Mastitis (400.0/0.0)

The attribute Swelling_Udder is Yes or No. It tells us about whether the cattle breast is swollen or not. If the breast is swollen, its value is Yes; otherwise its value is No.

The attribute temperature is High, Normal or Low. It tells us about whether the cattle body temperature is increased from the normal temperature or not. If the body temperature is greater than 39 values is High, the body temperature between 38 and 39 value is Normal; otherwise its value is Low.

Hence, for a certain disease to be classified as Mastitis both the antecedent of the rule ((swelling_udder) and (Temprature)) should be true. In other words, if the diseased cattle breast is swelling second, then that the cattle is affected by mastitis disease. If either of them is false, then the conclusion (class =Mastitis) will be false.

But Prolog does not work in IF… THEN format, rather it works in reverse order. PROLOG starts with a goal and then goes to the facts which can proof the goal as true. Therefore, the above rule has to be formatted as:

disease (mastitis): - (swelling_udder = Yes), (Temprature= High).

As illustrated above, the conclusion comes first with predicate =’class’ and followed by ‘**: -**‘ replacing ‘=>‘in JRip and then antecedents joined by ‘,’ replacing ‘and ‘in JRip rule. And finally Prolog rules terminate by period (.) whereas JRip rules terminate with new line.

7.4. Evaluation of the Prototype

System evaluation is the basic issue for the application of successful and effective knowledge base system. The developed system, cattle disease diagnosis and treatment is tested and evaluated to check whether the objectives of the research are achieved or not. The other one is preparing rules and feed the proposed system with these rules and give the same rule for domain experts and compare the results of the proposed system and the domain experts such that we can make sure that the proposed system could replace the domain experts in his or her absence. In addition to this, evaluation can be done by conducting a user acceptance testing which will help the researcher to make sure whether the proposed system is user friendly or not.

7.4.1. User Acceptance Testing

In this way, three users were selected and had been given the chance to use and interact with the system. The user selected were one Veterinarian Medicine Doctors (VMD) and two animal health assistant professional from Debre Birhan animal health center.

The user acceptance testing ensures how the users or domain experts view the system on the bases of the criteria‘s. Different researchers used different types of user acceptance testing, evaluation criteria [7]. The evaluators were allowed to rate the options as excellent, very good, good, fair and poor for those closed ended questions.

The table 7.2 below indicates the feedbacks obtained from the domain experts (evaluators) on system interactions as calculated based on the given scale.

Table 7.2 Performance Evaluation through visual interaction

No.	Criteria of Evaluation	Poor	Fair	Good	Very Good	Excellent	Average
1	Simplicity to use and interact with the prototype				2	1	4.33
2	Attractiveness of the prototype system			1	2		3.67
3	Efficiency in time			1	1	1	4
4	The ability of the prototype system in making right treatment and control of an identified disease				3		5
5	The accuracy of the prototype system in diagnosis of diseases of cattle				2	1	4.33
6	Knowledge adequacy regarding disease diagnosis in the Knowledge Based System			1	1	1	4
7	How do you rate the significance of the system in the domain area?				1	2	4.67

As table 7.2 points out, 33.3 % of the respondents reply easiness use and to interact with the prototype knowledge base system as **excellent** the highest number (66.7%) rated easiness to use and to interact with the prototype knowledge base system as **very good**. In case of attractiveness of the prototype, 33.33% of the respondents reply the prototype knowledge base system as **good**, and the highest number which is 66.7% rated the attractiveness of the prototype knowledge base system as **very good**. In case of time, efficiency, the same number of evaluators (33.33%) rated the prototype knowledge base system as **good**, **very good** and **excellent**. Regarding The ability of the prototype system in making right treatment and control of an identified disease, 100% of respondents reply the prototype knowledge based system as **very good**. Moreover, 66.7% of the respondents evaluate the prototype knowledge base system as **good** and 33.3% **excellent** at providing accurate for diagnosis of diseases of cattle. For criteria of the prototype incorporating sufficient knowledge, the same number of evaluators (33.33%) rated the prototype knowledge base system as **good**, **very good** and **excellent**. Finally, concerning the significance, diagnosis knowledge based system, 33.3% of the evaluated, responded as **very good** and 66.7% responded as **excellent**.

The reply from cattle disease diagnosis and treatment domain experts, who used the system, is that the system is appropriate, applicable and capable. Also results obtained from domain experts during interaction with the system and closed ended question, all of the evaluators responded the system as good, very good and excellent and none of them respond poor and fair for any section of the knowledge based system. Table 7.3 indicates the number of responses obtained for each of the options that the respondents rate to evaluate the prototype.

Table 7.3 Experts Response in Closed ended Question

Evaluators who respond as	Value	Total number of responses for each option of all seven questions	Percentage
Poor	1	0	0.00%
Fair	2	0	0.00%
Good	3	3	14.29%
Very Good	4	12	57.14%
Excellent	5	6	28.57%

As shown in the above table 7.3, based on closed ended questions as evaluation criteria, the domain experts reply the prototype as very good twelve times (57.14%). The experts also respond the prototype as excellent six times (28.57%). The least response is good three times (14.29%) regarding the prototype on its easiness, incorporation of sufficient knowledge.

As table 7.2 reveals, the total average performance the prototype from the cattle disease diagnosis and treatment domain expert evaluation is 4.29 out of the three selected experts on the scale 5=excellent, 4=very good, 3=good, 2= fair and 1=poor. Therefore, the total average performance of the prototype is found to be 85.8%. But, this measure alone is not enough to measure performance of the prototype since it tells us the overall performance for the next section 7.4.2. Generally, this prototype has got very good acceptance of the domain experts in predicting the disease and providing treatment for each disease. With time efficiency and cost effectiveness, this prototype disease diagnosis and treatment KBS has great value in predicting disease of cattle diagnosis and treatment by using knowledge base stored as rules and facts of disease diagnosis and treatment is needed in the areas it can be implemented.

7.4.2. System Performance Testing using Test Cases

System performance testing is done by preparing test cases. The test cases include samples of disease instances taken from BAHC dataset. The instances include 20 attributes with their respective values. The test cases, which are unlabeled disease instances are delivered by domain experts to label them as Blackleg, Anthrax, FMD, Pneumonia, LSD, Mastitis, Enteritis and Internal parasite. Considering the numbers of attributes and the time it consumes to label it manually, the researcher prepared only 40 test cases/instances for system performance testing.

Confusion matrix is used for comparing the performance of diagnosis and treatment of cattle disease KBS with domain expert's judgment. System performance testing basically used to measure Precision, Recall, F-measure, and True Positive rate how accurate the system is.

Table 7.4 Confusion matrix for evaluate experts 'judgment to experts 'judgment

	Performance of Combined rule based system									
	Class	FMD	Blackleg	Enteritis	LSD	Pneumonia	Internal Parasite	Anthrax	Mastitis	Total
Domain experts feedback	FMD	1	0	0	0	0	0	0	0	1
	Blackleg	0	6	0	2	0	0	0	0	8
	Enteritis	0	0	4	1	0	0	0	0	5
	LSD	0	0	0	2	0	0	0	0	2
	Pneumonia	0	0	0	0	12	0	1	0	13
	Internal parasite	0	0	0	0	0	9	0	0	9
	Anthrax	0	0	0	0	0	0	1	0	1
	Mastitis	0	0	0	0	0	0	0	1	1
	Total	1	6	4	5	12	9	2	1	40

The confusion matrix on table 7.4 shows the matrix of test case evaluation by diagnosis and treatment KBS and domain experts' judgment. The rows illustrate evaluation of domain expert and the columns illustrate the result of diagnosis and treatment KBS. The entries under column FMD testified that out of one instance is correctly identified as FMD disease.

The entries for Blackleg and Enteritis columns show that the system has correctly identified six instances are as blackleg, four instances are as Enteritis disease types respectively. The entries under column LSD testified that out of five instances, two of the instances are correctly identified as LSD disease and two instances are incorrectly identified as Blackleg and also one instance is incorrectly identified as Enteritis.

The entries for Pneumonia and Internal Parasite columns show that the system has correctly identified twelve instances are as Pneumonia and nine instances as internal parasite disease types respectively. The entries in the confusion matrix under Anthrax column depict that one instance out of two is incorrectly identified by the system as Pneumonia, one instance is correctly classified as Anthrax. Finally, the entries under column Mastitis testified that out of one instance is correctly identified as Mastitis disease.

Identification of each class of disease to their correct class is important to provide proper diagnosis and treatment to veterinaries so that they can take appropriate measures. But as shown in the confusion matrix, 2 blackleg are incorrectly identified as LSD disease, 1 Enteritis is incorrectly identified as LSD and 1 Pneumonia is incorrectly identified as Anthrax disease.

The system has correctly diagnosis or identified 36 test instances out of 40 to their correct class. This means the system has 90% diagnostic accuracy and 4 instances out of 40 are incorrectly classified which is 10%.

But, this measure alone is not enough to measure performance of the system since it only tells us the overall performance. Hence, Precision and Recall are employed to evaluate the system performance apart from diagnosis accuracy. Recall is the proportion of real positive cases that are correctly predicted positive. Precision denotes the proportion of predicted positive cases which are correctly real positives.

As clearly illustrated in table 7.5, the system‘s performance is evaluated in term of TP rate, Precision, Recall and F-measure, which enables us to view in detail how accurate, is the system is identifying cattle disease.

Table 7.5 TP Rate, Precision, Recall and F-Measure Evaluation Result in JRip

	TP Rate	Precision	Recall	F-Measure	Class
	99.8%	98.5%	99.8%	99.1%	FMD
	97.4%	99.6%	97.4%	98.5%	Blackleg
	98.7%	97.2%	98.7%	97.9%	Enteritis
	99.6%	96.5%	99.6%	98.0%	LSD
	96.4%	99.5%	96.4%	97.9%	Pneumonia
	98.0%	98.9%	98.0%	98.5%	Internal Parasite
	99.8%	99.6%	99.8%	99.7%	Anthrax
	99.8%	100%	99.8%	99.9%	Mastitis
Weighted Ave	98.7%	98.7%	98.7%	98.7%	

The system has registered its highest 99.8% Recall for FMD, Anthrax and Mastitis disease, and 99.6 for LSD, 98.7% for enteritis, 98.0% for Internal Parasite, 97.4% for Blackleg, 96.4% for Pneumonia disease. The system, registered its highest 100% Precision for Mastitis. The system has 99.6% Precision for Blackleg and Anthrax, 99.5% for Pneumonia, 98.9% for Internal Parasite, 98.5% for FMD, 97.2% for Enteritis and 96.5% for LSD. As compared to Precision (98.7%), the system has performed equal in its Recall (98.7%) for identifying diagnosing cattle disease accurately

7.5. Discussion

At the beginning, this study has four research questions to answer and let us discuss how these Questions have been answered by this study. The first research question of this study was ‘What are the main attributes that can properly predict the type of cattle disease’? To answer this question, information gain method and the domain expert’s interview were used and this study finds out that all instances of the rule generated discloses that the strong and significant attribute for predicting the cattle disease performance is a swell of neck.

The second question was “Which classification algorithm is best to develop the prediction model that can predict the type that will give to the diseased cattle which is infected by the disease? To answer this question, two experiments for three classification algorithms namely Naïve Bayes, J48 pruned, and JRip under 10-fold Cross-Validation test option/mode and percentage split were conducted and the experiments showed that JRip classification algorithm is the best classification algorithm to develop the prediction model that can predict the type of diagnosis that should be given to the diseased cattle is infected by disease because it registers better performance with 98.68% evaluation result and the researcher decided to use the results for further use in the development of knowledge base of KBS.

The third question was “how to acquire, model, represent and implement a data mining result with knowledge based system for the diagnosis and treatment of selected disease? To answer this question, a prototype Knowledge Based were developed using the knowledge that is acquired from domain Experts’ Interview and documents analysis with Prolog for treatment purpose and constructing disease diagnosis model, data mining technology is used to acquire hidden knowledge about the behavior of different disease from the given dataset. WEKA was integrated with Java by adding the WEKA jar file in Java library which enabled us to integrate the WEKA result automatically with the knowledge base. Then to call the knowledge base that is constructed with Prolog from Java, we have added JPL and jar file in Java library and the evaluation results of system performance testing and user acceptance testing showed that the proposed system registers better performance.

The final question was How to integrate data mining result with knowledge base system for developing the prototype? Java programming was used to integrate diagnosis of cattle disease model created using JRip rule with the Knowledge Based System automatically. For this study, we used Weka, JPL and swi-prolog tools to construct a predictive model by using SWIWEKA module. Hence, the system automatically takes the dataset, construct the model using the selected JRip rule classification algorithm and integrate the model with knowledge base for diagnosis of cattle disease. Here table 7.6 illustrates the researcher tries to summarize the related works in the following table for clear review:

Table 7.6 Comparison of different researchers in this study

Author	Title	Techniques	Evaluation Criteria's	
			User acceptance Testing	System performance Testing
Derejaw Lake	Developed on Web-based expert system for diagnosing cattle diseases	Interviews, document analysis	87.2%	___
Ahmad et al	Develop web based cattle disease expert system diagnosis with forward chaining method in Indonesia	Interviews	___	___
Oluwatoyin and Adedoyin	Developed a fuzzy expert system for diagnosis of cattle disease in Nigeria.	Observation & Interview	___	___
Rong Daoliang	Developed a web based expert system for diagnosis of cow disease	Interview	___	___
Birhane Bekele	Developed a prototype KBS for diagnosis and treatment of diseases of sheep and goats	Interview & Document analysis	84.8%	___
Tesfamariam Mulugeta	Integrating Data Mining Results with the Knowledge Based System for Diagnosis and Treatment of Visceral Leishmaniosis	Interview, Document Analysis, PART,J48 & JRip	86%	95%
Tadesse Beyene	Integrating Data Mining Results with Knowledge Based System for Cattle Disease Diagnosis and Treatment	Interview, Document Analysis, J48,JRip & Naïve Bayes	85.8%	90%

CHAPTER EIGHT

CONCLUSION AND RECOMMENDATION

8.1. Conclusion

Ethiopia is one among the nations that possesses the largest livestock population in the African continent with an estimated fifty six Million of cattle, fifty eight Million of sheep and goats and ten Million of equines, one Million of camels and fifty seven Million of chicken [1]. Ethiopia has great potential for increasing livestock production, both for local use and export. However, development has been constrained by numerous reasons. Cattle disease is the main constraint.

In this study, the possibility of integrating data mining result with knowledge based system is realized and explored. The integration process begun by taking samples of BAHC dataset. The dataset is preprocessed and made suitable for mining steps. Due to several limitations in acquiring knowledge for knowledge base from domain experts in the area of diagnosis and treatment of cattle disease, an automatic knowledge acquisition mechanism is proposed in this study. Data mining has proven to induce hidden knowledge from large collections of datasets. Hence, data mining classifier, JRip is employed for knowledge acquisition step since it has performed best among the selected classifiers with an accuracy of 98.68%.

To identify the best prediction model for diagnosis and treatment of cattle disease, 6 experiments for three classification algorithms, namely J48 pruned, Naïve Bayes and JRip under ten-fold Cross-Validation test option and percentage split test option were conducted. Finally, by conducting objective and subjective interestingness measure, the researcher decided to use rules that are generated by JRip classification algorithm model for further use in the development of knowledge base system because it registered better performance than J48 and Naïve Bayes with 98.68%, 98.65% and 98.48% evaluation result in 10-fold cross validation respectively.

The prototype Knowledge Based System, which provides advice for Animal Health Workers about diagnosis and treatment of cattle disease was developed using SWI-Prolog 7.7.13 with NetBeans 8.2. The proposed Knowledge Based System has Knowledge Base, Inference Engines, Explanation Facility and User Interface. Then 40 test cases were prepared to evaluate the performance of the proposed system. Finally, system performance evaluation, testing and user acceptance testing were conducted. User acceptance testing is performed based on seven criteria of evaluation. Selected domain experts are trained and used the system to evaluate how much the KBS meets their requirements. The system on average scored 85.8% based on user acceptance evaluation. The system has registered 98.7% overall accuracy according to the recall and precision.

However, further exploration and study has to be done to refine and yield a better Knowledge based system which can be deployed in real animal agriculture and to provide advice veterinaries so that they can take timely and appropriate actions for a certain cattle disease diagnosis and treatment.

Moreover, this study has covered the way for local researchers on using automatic knowledge acquisition techniques for the development of knowledge based system and motivates them to apply this approach than the conventional knowledge acquisition approach.

8.2. Recommendation

In this study promising result is achieved in integrating machine learning induced patterns with knowledge based system for diagnosis and treatment of cattle disease and providing advice to animal health workers. Some challenges have been encountered which hinder the system of scoring a better achievement.

The first one is in the course of integration, two interfaces, namely; graphical user Interfaces (for the Integrator) and command line interface for diagnosis and treatment of cattle disease KBS has been used. A challenge has encountered in bringing the Integrator and diagnosis and treatment KBS together under one interface. This is reflected on user acceptance test in that evaluator rated the simplicity to use and interact with the system below very good. The other challenge encountered is about using knowledge which the KBS has already used previously before re-running the Integrator application following a change in the numbers of the dataset. In addition, the designed prototype KBS supports eight types of disease, namely FMD, Blackleg, Enteritis, LSD, Pneumonia, Internal Parasite, Anthrax and Mastitis.

Hence the researcher believes further researches have to be done to boost the benefits of Integration of data mining with the knowledge based system and the following are recommended for future study:

- Building hybrid knowledge based system which is capable of employing rule based reasoning and case based reasoning with integrated data mining techniques.
- Improve and build knowledge base system with graphical user interface which makes simple to use and attractive to the user. Using different programming languages like Java, C# and so… on.
- Further combinations such as artificial intelligence, soft computing and other clustering algorithms can be used to improve the diagnosis accuracy and to reduce the rate of falls a negative rate and false positive rate.
- Integrating Text Mining Results with the Knowledge Based System. While we did the data mining process, we understood that there are notes that have been written by health experts as a remark. So that, if someone can conduct text mining and generate some case, it will expand the knowledge base and improve the performance of the system.
- It is known that most of Ethiopian agricultural areas have no facilities like electricity and a computer access. So, further research must be conducted to take the integrated knowledge based system in to mobile application.

References

[1] B. Akalu, "Review on Epidemiology of Bovine Tuberculosis in Ethiopia," *Academic Journal of Animal Diseases* , vol. 6, no. 3, pp. 56-66, 2017.

[2] "Physiology of Cattle," [Online]. Available: https://www.farmsanctuary.org/wp-content/uploads/2012/06/Animal-Care-Cattle.pdf. [Accessed 28 02 2018].

[3] S. Dr. Ashok, "Common Cattle Diseases: Symptoms, Treatment And Prevention," vol. 453, no. 446, pp. 77-85, 2014.

[4] *The Drug Administration and Control Authority of Ethiopia, Standard Veterinary Treatment Guidelines for veterinary clinics*, 2006.

[5] A. Rajendra, P. Akerkar and S. Srinivas, "Knowledge-Based systems," Jones and Bartlett, Sudbury: MA 01776, 2010.

[6] p. Sajja and R. Akerkar, "Advanced Knowledge Based Systems," *Model, Applications and Research*, vol. 1, p. 1 – 11, 2010.

[7] A. Mohammed, "Towards Integrating Data Mining with Knowledge Based System: The Case of Network Intrusion detection," M.Sc. Thesis, Adis ababa University, 2013.

[8] M. Tefera, "Global Crisis and the Challenge of Veterinary Teaching in Ethiopia," *Global Veternarian 5*, vol. 6, pp. 294-30, 2010.

[9] L. Derejaw, "Web-based expert system for diagnosing cattle diseases," M.Sc. Thesis in Computer Science, Addis Ababa University, Addis Ababa, Ethiopia, 2013.

[10] NetBeans, 06/04/2015. [Online]. Available: https://netbeans.org/features/.

[11] [Online]. Available: https://en.wikipedia.org/wiki/Cattle. [Accessed 05 01 2018].

[12] K. Mehmed and B. John, Data mining concepts, Models, Methodes, and Algorithms, USA: Wiley -IEEE Press, 2003.

[13] U. Fayyad, G. Piatetsky-Shapiro and P. & Smyth, "From Data mining to knowledge Discovery," 1996. [Online]. Available: http://citeseer.nj.nec.com:fayyad96from.htm. [Accessed 10 02 2018].

[14] M. C. Azevedo, Data Mining Descoberta de Conhecimento em bBase de Dados, FCA, 2005.

[15] S. Institute, *SAS Enterprise Minor SEMMA,* 2016.

[16] K. Charly, "Data Mining for the Enterprise," *31st Annual Hawaii Int. Conf. on System Sciences,* vol. 7, pp. 295-304, 1998.

[17] K. J. Cios, P. Witold, S. Roman W. and K. Lukasz A., Data Mining: A Knowledge Discovery Approach, USA: Springer Science & Business Media, 2007.

[18] G. Awoke, "Predicting HIV Infection Risk Factor using Volantory Counseling and Testing Data," Addia Ababa University , Addis Ababa, Ethiopa, 2012.

[19] A. B. Michael J. and S. L. Gordon, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, Second ed., Indianapolis, Indiana: Wiley Publishing, Inc., 2004.

[20] E. W. T. Ngai, L. Xiu and D. C. K. Chau, "Application of data mining techniques in customer relationship management," *A literature review and classification. Expert Systems with Applications,* vol. 36, pp. 2592-2602 , 2009.

[21] G. S.D., K. A.S. and T. S.M., "Data Mining: Task, Tools, Techniques and Application," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 3, no. 10, pp. 8095-8098, 2014.

[22] C. G. Carrier and O. Povel, "Characterising data mining software," *Intelligent Data Analysis*, vol. 7, pp. 181 - 192., 2003.

[23] S. V.Sathiya and N. Dr.Sai Satya, "Data Mining Tasks Performed By Temporal Sequential Pattern," *International Journal of Research and Computational Technology,* vol. 2, no. 3, pp. 1-6, 2012.

[24] H. Jiawei, K. Micheline and P. Jian, Data Mining Concepts and Techniques, Third ed., New York: Morgan Kaufmann Publishers is an imprint of Elsevier, 2012.

[25] K. Dorina, "Student Performance Prediction by Using Data Mining Classification Algorithms," *International Journal of Computer Science and Management Research ,* vol. 1, no. 4, pp. 686-690, 2012.

[26] M. Vincent, "Introducing a data mining process framework to enable consultants to determine effective data analytics tasks," Master Thesis, University of Technology, Delft , 2012.

[27] H. Tesfahun, "Application of Data Mining for Predicting Adult," MSC Thesis, Addis Ababa University, Addis Ababa, 2012.

[28] C. d. R. Bruno and T. d. S. J. Rafael, "Identifying Bank Frouds using CRISP-DM and Decision Tree," *International journal of computer science & information Technology (IJCSIT),* vol. 2, no. 5, pp. 162-169, 2010.

[29] P. Thair Nu, "Survey of Classification Techniques in Data mining," *Survey of Classification Techniques in Data mining,* vol. 18, no. 20, pp. 978-988., 2009.

[30] C. William W., "Fast Effective Rule Induction ," in Machine Learning," *To appear in \Machine Learning: Proceedings of the 12th International Conference (ML95),* pp. 115-123, 1995.

[31] B. S. Adriane and C. B. Antonio, "Knowlede Discovery for Classification of Three-Phase Vertical Flow Patterns of Heavy oil from Pressure Drop and Flow Rate Data,"," *Jornal of Petroleum Engineering , vol. 2013, pp. 1-8, August 2010.,* vol. 20, no. 13, pp. 1-8, 2010.

[32] P. Tina R. and S. Mrs. S. S., "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *International Journal Of Computer Science And Applications* , vol. 6, no. 2, pp. 256-261, 2013.

[33] B. D. Mahamana, H. C. Passamb, A. B. Sideridisb and C. P. Yialouris, "DIARES-IPM: a diagnostic advisory rule-based expert system for integrated pest management in Solanaceous crop systems," *Agricultural Systems,* vol. 76, pp. 1119-1135, 2003.

[34] R. Engelmore and E. Feigenbaum, "Advanced Research in Knowledge-based Systems: Inventing the next generation," 2003. [Online]. Available: http://www.wtec.org/loyola/kb/c1_s5.htm. [Accessed 31 01 2018].

[35] R. Raman and K. Prasad, "Applications of Knowledge Based Systems in Mining," *APCOM 87. Proceedings of the Twentieth International Symposium on the Application of Computers and Mathematics in the Mineral Industries,* Vols. 167- 180, pp. 1-11, 1987.

[36] K. P. Tripathi, "A Review on Knowledge-based Expert System:Concept and Architecture," *IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications",* pp. 19-23, 2011.

[37] J. Mare and Vlaanderen, "Automated Knowledge Acquisition for Expert Systems an Overview," PhD dissertation, Faculty of Philosophy, Erasmus University Rotterdam, Netherlands, 1990.

[38] M. Prof. and P. OPREA, "On the Use of Data-Mining Techniques in Knowledge-Based Systems," *Department of Informatics, University Petroleum-Gas of Ploieşti,* vol. 1, no. 4, pp. 21-24, 2006.

[39] A. Abraham, "Rule-based Expert Systems. Handbook of measuring system design," pp. 909-919, 2005.

[40] B. K. Birhane, "Developing Knowledge Based System for Diagnosis and Treatment of Diseases of Sheep and Goats," Msc Thesis, Debre birhan, 2015.

[41] G. Stephan, H. Pascal and A. Andreas, "Knowledge Representation and Ontologies Logic, Ontologies and Semantic Web Languages," FZI Research Center for Information Technologies, University of Karlsruhe, Germany and Institute AIFB, University of Karlsruhe, Germany, [Online]. Available: daselab.cs.wright.edu/resources/publications/kr-onto-07.pdf. [Accessed 15 12 2017].

[42] R. Stuart and N. Peter, Artificial Intelligence A Modern Approach, 3rd ed., Prentice Hall: Engle wood Cliffs, New Jersey 07632, 2010.

[43] H. Vasant, "Artificial intelligence: An overview," Iowa State University, [Online]. Available: https://pdfs.semanticscholar.org/63c0/5f0e837e2e6d132bf3ccd5f79d9500cf1ebc.pd. [Accessed 15 12 2017].

[44] B. Max, *Principles of Data Mining*, UK: Portsmouth Springer, 2007.

[45] Z. Ahmad, S. Wenni and S. Muhamad, "Web Based Cattle Disease Expert System Diagnosis with forward Chaining Method," *International Conference on Environment and Technology (IC-Tech)*, pp. 1-8, 2017.

[46] O. Agbonifo and A. Ajayi, "Design of a Fuzzy Expert Based System for Diagnosis of Cattle Diseases," *International Journal of Computer Applications & Information Technology*, vol. I, no. III, pp. 93-98, 2012.

[47] R. Libin and L. Daoliang, " A Web Based Expert System for Milch Cow Disease Diagnosis System in China," *in IFIP International Federation for Information Processing,* vol. 259, p. 1441–1445, 2008.

[48] M. Tesfamariam, "Integrating Data Mining Results with the Knowledge Based System for Diagnosis and Treatment of Visceral Leishmaniasis," MSC Thesis, Gondar University, Gondar, 2015.

[49] G. Solomon, "A Self-Learning Knowledge Based System for Diagnosis and Treatment Of Diabetes," Addis Ababa, 2013.

[50] A. Gebremedhin, "Major Animal Health Problems of Market Oriented Livestock, Development in Atsbi Womberta Woreda, Tigray Regional State," DVM thesis in Veterinary, Medicine, Addis Ababa University, Debre Zeit, Ethiopia, 2007.

[51] T. Tadesse, M. Mulisa and F. Teka, "Bovine Mastitis: Prevalence and Isolation of Major Pathogens in Dairy Farms of Selected Sites in Addis Ababa, Ethiopia," *Applied Journal of Hygiene,* vol. 3, no. 3, pp. 31-37, 2014.

[52] S. Hunderra, Z. Ademe and A. Sintayehu, "Dairy Cattle Mastitis In and Around Sebeta, Ethiopia," *Intern J Appl Res Vet Med,* vol. 3, no. 4, pp. 332-338, 2005.

[53] [Online]. Available: http://kb.rspca.org.au/What-is-mastitis-in-dairy-cows_221.html. [Accessed 2 2 2017].

[54] A. Zelalem, D. Hailu and G. Getachew, "Assessment of Distribution and Associated Risk Factors of Lumpy Skin Disease in Selected Districts of West Wollega Zone, Western Ethiopia," *Academic Journal of Animal Diseases,* vol. 4, no. 3, pp. 130-140, 2015.

[55] S. Astaws, T. Alebachew, B. Eskeziaw and G. Abebaw, "Major Health Challenges of Dairy Cattle in and Around Wukro, Ethiopia," *World Journal of Dairy & Food Sciences,* vol. 11, no. 2, pp. 156-162, 2016.

[56] M. Boris and M. Milan, "Prediction and Decision Making in Health Care using Data," *International Journal of Public Health Science (IJPHS)*, vol. 1, no. 2, pp. 69-78, 2012.

[57] Y. Chen, Learning classifiers from imbalanced, only positive and unlabeled data Sets, USA: Department of Computer Science, Iowa State University, 2009.

[58] L. Wenyuan and G. K. Janbin, "Integrating Artificial Intelligence and Data Mining," *Integrating Artificial Intelligence and Data, Gold Coast.*, vol. 84, pp. 99-107 , 2007.

[59] U. Endriss, An Introduction to Prolog Programming, 2 ed., Netherland: Universiteit van Amsterdam, 2013.

[60] D. Merritt, Building Expert Systems in Prolog, USA: Amzi! inc, 2000.

[61] "SWI-Prolog Tutorial," [Online]. Available: www.swi-prolog.org/packages/JPL. [Accessed 10 06 2018].

[62] K. Dr. D. Ashok and G. R., "Performance and Evaluation of Classification Data Mining Techniques in Diabetes," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1312-1319, 2015.

[63] C. Alebachew and M. Million, "An Intelligent Network Intrusion Detcetion System Using Data Mining and Knowledge Based System," *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 17, pp. 4273-4282, 2017.

Appendix I

Interview questions

The importance of this interview questions is to extract tacit knowledge from experts in disease of cattle diagnosis, which help for the development of Knowledge Based System for disease diagnosis and treatment**.** The interviewer writes the answerers replied using pen and notebook. I would like to thank for your collaboration and precious information.

1. Which types of cattle diseases are common in Ethiopia?
2. How a disease is identified when cattle is infected by a certain disease?
3. What are the overall symptoms that are shown when the diseases are occurrence?
4. What are the steps to diagnose a certain disease?
5. How can treat and control the cattle diseases?
6. What are the average range of the age of calf, young and adult in both breeds?

Appendix II

User acceptance Testing evaluation criteria

Dear Evaluator,

This evaluation form is prepared aiming at measuring to what extend does diagnosis and treatment of cattle disease KBS is useable and acceptable by end users in the area of Animal health center. Therefore, you are kindly requested to evaluate the system by labeling (√) symbol on the space provided for the corresponding attribute values for each criteria of evaluation.

I would like to appreciate your collaboration in providing the information.

Note:- the values for all attributes in the table are rated as: Excellent=5, Very good =4, Good=3, Fair= 2 and Poor =1.

No.	Criteria of Evaluation	Poor	Fair	Good	Very Good	Excellent	Average
1	Simplesity to use and interact with the prototype						
2	Attractiveness of the prototype system						
3	Efficiency in time						
4	The ability of the prototype system in making right treatment and control of an identified disease						
5	The accuracy of the prototype system in diagnosis of diseases of cattle						
6	Knowledge adequacy regarding disease diagnosis in the Knowledge Based System						
7	How do you rate the significance of the system in the domain area?						

Appendix III

Initial list of original attributes with their description

Attribute name	Description	Datatype	Domain value
Date	Diseased cattle date of registration	Date/Time	{Yy/mm/DD format}
Kebele	Address of the cattle	Numeric	{Continuous value}
Sex	Cattle gender	Nominal	{Male, Female}
Spp	Cattle species	Nominal	{Bovine}
Breed	Cattle offspring	Nominal	{Local, Cross}
Age	The age of the cattle	Nominal	{Adult, young, calf}
Temperature	The rectal temperature of the cattle	Numeric	{Continuous}
Fever	Cattle body temperature that is higher than normal	Binary	{Yes, No}
Cough	To force air through your throat with a short	Binary	{Yes, No}
Painful	Causing pain to cattle body	Binary	{Yes, No}
Diarrhea	Cattle intestine evacuation with More or less fluid stools	Binary	{Yes, No}
Nasal discharge	Mucus flows out of cattle nose	Binary	{Yes, No}
Abortion	Cows/ Heifers embryo or fetus are death	Binary	{Yes, No}
Loss of appetite	Cattle decrease feeding	Binary	{Yes, No}
Swelling of neck	Cattle neck swells	Binary	{Yes, No}
Lameness	Cattle impair freedom of movement	Binary	{Yes, No}
Crepitation sound	Cattle make a sound of crackling sound	Binary	{Yes, No}
Change color of milk	Normal milk color change	Binary	{Yes, No}
Decrease milk yield	Reduce the milk	Binary	{Yes, No}
Swelling udder	A bag like organ containing the cattle glands	Binary	{Yes, No}
Shivering	The cattle body Shake slightly	Binary	{Yes, No}

Sore feet	Cattle foot pain	Binary	{Yes, No}
Salivation	The cattle drool the mucus	Binary	{Yes, No}
Rolling	Revolving the cattle around the surface	Binary	{Yes, No}
Smacking lips	Involving a loud smacking of the lips	Binary	{Yes, No}
Depression	Decrease and loss of interest in pleasurable activities	Binary	{Yes, No}
Nodular lesion	The cattle body is inflamed	Binary	{Yes, No}
[1]Abnormal breathing	Fast breathing (depend on age) rate/minute	Binary	{Yes, No}
Rough hair	Standing the cattle hair	Binary	{Yes, No}
Emaciation	The cattle become very thin	Binary	{Yes, No}
Poor body condition	Failing of state of physical fitness	Binary	{Yes, No}
Diagnosis	Confirmative diagnosis of cattle disease	Nominal	{Blackleg, FMD, LSD, anthrax, mastitis, Enteritis, Pneumonia, Internal Parasite}

[1] Calf, newborn=56, 2 weeks=50, 5 weeks=37, 6 months=30, yearling= 27 and cattle adult=12-16 [4].40

Appendix IV

An integrated knowledge based system for diagnosis of cattle disease diagnosis

Sample prolog code

:-[Classifier]:-

Start:-

Write (‘Select the Value of Temperature: From The option’→’),read (temp),

Write (‘Select the Value of Smacking lips: From The option’→’),read (smacking),

Write (‘Select the Value of Sore Feet: From The option’→’),read (sore),

Write (‘Select the Value of Fever: From The option’→’),read (fever),

Write (‘Select the Value of Lameness: From The option’→’),read (lame),

Write (‘Select the Value of Depression: From The option’→’), read (depression),

Write (‘Select the Value of Painful: From The option’→’), read (painful),

Write (‘Select the Value of Crepitation Sound: From The option’→’), read (sound),

Write (‘Select the Value of Loss of Appetite: From The option’→’),read (appetite),

Write (‘Select the Value of Diarrhea: From The option’→’),read (diarrhea),

Write (‘Select the Value of Nodular lesion: From The option’→’),read (lesion),

Write (‘Select the Value of Nasal discharge: From The option’→’),read (discharge),

Write (‘Select the Value of Coughing: From The option’→’),read (cough),

Write (‘Select the Value of Abnormal Breath: From The option’→’),read (breath),

Write (‘Select the Value of Swell of Udder: From The option’→’),read (udder),

Write (‘Select the Value of Decrease Milk Yield: From The option’→’),read (milkyield),

Write (‘Select the Value of Change Color of Milk: From The option’→’),read (colormilk),

Write (‘Select the Value of Swell of Neck: From The option’→’),read (neck),

Write (‘Select the Value of Rough Hair: From The option’→’),read (hair), Write (‘Select the Value of Poor Body Condition: From The option’→’),read (condition),

Jpl_call(‘writetotext.Write To Text’, writeToSample,[Temperature,Smacking_lips,Sore_Feet,Fever,Lameness,Depression,Painful,Crepetation_sound,Loss_of_Appetite,Diarrhea,Nodular_lesion,Nasal_discharge,Coughing,Abnormal_breath,Swell_of_udder,Decrease_milk_yield,Change_color_of_milk,Swell_of_neck,Rough_hair,Poor_body_condition],_),

wkpl_read_arff(‘Tadeset.arff’, Mainarff),

wkpl_set_classIndex(Mainarff,20),

wkpl_classifier(‘weka.classifiers.rules.JRip’, Classifier),

wkpl_build_classifier(Mainarff, Classifier),

decrement (Number, Index),

Wkpl_get_instance(dataset, Index, Instoclassify),

Wkpl_classify_instance(Instoclassify, Classifier, Prediction),

Write(Prediction),nl,

Pred_value(Prediction).

Pred_value(Prediction):-

(Prediction==0.0) →write (‘FMD’);

(Prediction==1.0) →write (‘Blackleg’);

(Prediction==2.0) →write (‘Enteritis’);

(Prediction==3.0) →write (‘LSD’);

(Prediction==4.0) →write (‘Pneumonia’);

(Prediction==5.0) →write (‘Internal Parasite’);

(Prediction==6.0) →write (‘Anthrax’);

(Prediction==7.0) →write (‘Mastitis’), nl.