ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

School of Electrical Engineering and Computing

Department of Computing

Integrated Predictive Model and Knowledge Based System for Wheat Disease Detection: Case of Arsi Zone

By: - Fayisa Dekebi Tusamo

Adama, Oromiya, Ethiopia

January 2018

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

School of Electrical Engineering and Computing

Department of Computing

Integrated Predictive Model and Knowledge Based System for Wheat Disease Detection: Case of Arsi Zone

A Thesis Submitted to the Department of Computing at Adama Science and Technology University in Partial Fulfillment of the Requirements for Degree of Master of Science in Information Systems

By: - Fayisa Dekebi Tusamo

Advisor: - Tilahun Melak [PhD]

Adama, Oromiya, Ethiopia

January 2018

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

School of Electrical Engineering and Computing

Department of Computing

Integrated Predictive Model and Knowledge Based System for Wheat Disease Detection: Case of Arsi Zone

By: - Fayisa Dekebi Tusamo

Name and Signature of the Examining Board Members

_____	_____
Advisor	Date and Signature
_____	_____
Chairperson of examining board	Date and Signature
_____	_____
Internal examiner	Date and Signature
_____	_____
External examiner	Date and Signature

DECLARATION

I declared that this thesis is my original work and has not been submitted as a partial requirement for any degree in any university and that all source and materials used for this thesis have been duly acknowledged.

Name: Fayisa Dekebi Tusamo

Signature: _____

Date: _____

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

Tilahun Melak [PhD]

Signature: _____

Date: _____

ACKNOWLEDGEMENTS

First of all, I would like to be glad to my almighty God for his unreserved provision and help during this research work. I am honestly thankful to my advisor, Dr. Tilahun Melak for his advice, guidance, constructive and on time comments while undertaking this study. I also forward my sincere gratitude to Kulumsa Agriculture Research Center and its staff for allowing undertaking this study and their endless support from collecting the experimental data for this study to general description and comments on the nature of the dataset. I want to thank Dr. Bekele Hunde the plant pathologist and researcher for his endless support and comment from beginning to the end of this study. Lastly, I would like to thank and appreciate my all my family and friends who encouraged and supported me in accomplishing this thesis.

Table of Contents

ACKNOWLEDGEMENTS ---------- iii

LISTS OF FIGURES ---------- vii

LISTS OF TABLES ----------viii

ACRONYMS AND ABBREVIATIONS---------- ix

ABSTRACT---------- x

CHAPTER ONE---------- 1

INTRODUCTION ---------- 1

1.1 Background of Study ---------- 1

1.2 Motivation to Study---------- 5

1.3 Statement of the Problem ---------- 5

1.4. Objectives of the Study---------- 8

1.5. Scope and Limitations of the Study ---------- 8

1.6. Significance of the Study---------- 9

1.7. Organization of the Thesis ----------11

CHAPTER TWO----------12

LITERATURE REVIEW ----------12

2.1. Wheat Crops in Ethiopia ----------12

2.2. Data Mining and Knowledge Base Systems ----------13

2.3. Knowledge Based Systems ----------22

2.4. Review of Related Works----------34

CHAPTER THREE----------37

METHODOLOGY----------37

3.1. Research Area/Setting and Period ----------37

3.2. Method of Study----------37
3.3. Data Collection and Knowledge Discovery ----------38
3.4. Knowledge Representation ----------39
3.5. System Development and Implementation Tools ----------40
3.5.1. Implementation Tools ----------40
3.6. Testing and Evaluation ----------41
CHAPTER FOUR----------42
EXPERIMENTATION AND ANALYSIS----------42
4.1. Wheat Disease Diagnosis and Treatment Process ----------42
4.2. Data Understanding and Selection----------44
4.3. Data Preprocessing----------49
4.4. Experimentation ----------52
4.4.1. Building Predictive Models----------52
4.4.2. Models Evaluation and Selection Methods----------54
4.4.3. Evaluating and Using the Rules----------58
4.5. System Design and Architecture ----------62
CHAPTER FIVE ----------66
IMPLEMENTATION AND EVALUATION ----------66
5.1. Implementations----------66
5.2. Disease Detection and Activity Flow ----------68
5.3. Testing and Evaluation ----------74
5.3.1. System Performance Testing ----------74
5.3.2. User Acceptance Testing----------75
CHAPTER SIX----------79
CONCLUSION AND RECOMMENDATIONS----------79
6. 1. Conclusion----------79

6.2. Recommendations----------80

REFERENCES ----------81

ANNEXIES ----------87

Appendix I----------87

Appendix II----------90

Appendix III----------92

Appendix IV----------93

Appendix V----------94

LISTS OF FIGURES

Figure 2-1 Knowledge discovery process models adopted [29] ----------16
Figure 2-2 KBS Architecture Adopted [47] ----------26
Figure 2-3 Development of Knowledge-Based System, adopted [46]. ----------27
Figure 4-1 Attribute Ranks----------51
Figure 4-2 Sample ARFF Files ----------52
Figure 4-3 Predicted Accuracy of Models by Cross-Validation----------57
Figure 4-4 Times Taken By Classifiers ----------57
Figure 4-5 Summarized Error Rates ----------58
Figure 4-6 Architectural design and Framework for Implementation ----------64
Figure 5-7 Functional Flow of Activities ----------69
Figure 5-8 Home Page User Interface ----------69
Figure 5-9 User Interface Showing List of Rules----------70
Figure 5-10 User Interface Showing Activated Rules ----------71
Figure 5-11 User Interface for Adding Facts ----------72
Figure 5-12 User Interface When Septoria Disease Detected----------72
Figure 5-13 User Interface When Stem rust Disease Detected ----------73
Figure 5-14 User Interface When No Any Disease Detected ----------73

LISTS OF TABLES

Table 4-1 Descriptions of Initial Data Set 45
Table 4-2 Descriptions of Selected Attributes 47
Table 4-3 Predictor Variables with Number of Distinct Values 48
Table 4-4 Distribution of Dependent Variable 49
Table 4-5 Percentages of Attributes with Missing Value 50
Table 4-6 Performance of Classifiers 54
Table 4-7 Stratified Cross-Validations of Classifiers 55
Table 4-8 True Positive, Precision, Recall and F-Measure of Classifiers with Classes 56
Table 4-9 Rules Generated From JRip 59
Table 5-10 System Performance Testing Confusion Matrix 75
Table 5-11 User Acceptance Testing and Users Feedback 76

ACRONYMS AND ABBREVIATIONS

ARFF: Attribute Relation File Format

CBS: Case Based Reasoning

CRISP-DM: Cross Industry Standard Process for Data Mining

CSA: Central Statistical Agency

CLI: Command Line Interface

CSV: Comma Separated Values

DA: Development Agents

DM: Data Mining

GUI: Graphical User Interface

ID3: Iterative Dichotomiser

KB: knowledge Base

KBS: Knowledge Base System

KDD: Knowledge Discovery in Database

SNNPR: Southern Nations Nationalities and Peoples Region

RBR: Rule Based Reasoning

RQ1: Research Question One

RQ2: Research Question Two

WEKA: Waikato Environment for Knowledge Analysis

WDC-KBS: Wheat Disease Classifier Knowledge Base System

ABSTRACT

Ethiopian economy is Agriculture dependent as 80% of the country's population reliant on agriculture. Wheat covers the country's 26% of annual cereal crop production and supplying nutritional benefits. The recent outbreak of cereal crop diseases resulted in the loss of productivity throughout the country. There is a need for efforts of all stakeholders in agriculture. Data mining and KBS come across with varieties approaches and application in solving agriculture problems. The main objective of the study is to investigate the construction of a predictive model to extract hidden knowledge and integrate with the Knowledge Base System for detecting Wheat diseases. This could help in the effort of enhancing decision making in the agriculture sector. For this purpose Mixed research design, hybrid knowledge acquisition techniques, prototyping system development method, rule based knowledge representation, WEKA with JDK 8, and Java libraries like JavaFX, JESS and Eclipse have been selected and implemented to achieve the objective of the study. To identify the best prediction model for detection of wheat diseases four experiments were undertaken with four respective classification algorithms. Objective evaluation standards were used to evaluate the prediction accuracy and performance of implemented classifiers. As a result, JRip classification algorithm selected and integrated in the development of KBS since, it registered better performance result of 99.61 % accuracy with Objective evaluation method. In addition to this, system performance and users' acceptance testing have been undertaken to evaluate the prototype KBS. Test cases and questionnaire were prepared to evaluate the performance and users' acceptance of the developed system prototype. As a result, the system can achieve in the absence of domain experts 80 % system performance evaluation result; indicating the KBS has the necessary knowledge for detection and treatment of wheat diseases which in turn implies that the result of study was applicable and operative. Besides to this, the proposed system achieves 86.3% of the users' acceptance which indicates it could be operational, if implemented. Since the study achieved better results, it can be applicable for stated objective in araea.

***Keywords**: Data Mining, Knowledge-Based System, Wheat Disease, WEKA*

CHAPTER ONE

INTRODUCTION

1.1 Background of Study

The rapid development of computers and computer technology has transformed the method of agriculture operation which leads toward the generation of huge amount of agricultural data. But, developing nations like Ethiopia which their economy is agriculture dependent, are not advanced in technology for supporting their agriculture. Many researches have been conducted either on innovating and/or adopting advances in agricultural technology by different scholars. But this research work addressed the problems of Wheat disease throughout research area. The research dealt the context of current Ethiopian agricultural situation where most of the activities were undertaken by the human experts. Agriculture is the backbone of Ethiopian economy as the 80% of country's population dependent on agricultural activities. The country's population is 94.351 million in 2014, the second most-populous country in Africa with least urbanized having 79.8% of the rural population as census of Central Statistical Agency in 2013.

The small-scale farmers produce 94% of the food crops and 98% of the coffee; large-scale private and state commercial farms produce just 6% of food crops and 2% of the coffee grown [1]. The Wheat crop feeds about 2.5 billion poor people as world's most important food crop in around 90 countries and is a crucial source of calories and protein. But the demand for Wheat currently exceeds the world's ability to produce and the reason behind is the reduction in the production of the Wheat because of increasing new kinds of fungus infections, like Wheat Yellow rusts, Stem rust, Leaf rust and Septoria that are killing crops [2]. As to Ethiopia, Wheat is the main cereals crop cultivated throughout the country. Accordingly, Ranked as third most produced grain after sorghum and maize in area coverage of (7.39%) and second in total production with (3.24%). It mainly produced as the main indispensable food for about 36% of the population (CSA, 2004).

The area covered under Wheat production is estimated to be about 1.5 million hectares, which makes the country the largest Wheat producer in Sub-Saharan Africa. The national average yield of Wheat in the country is 3.3 tons per hectare. The Wheat produced in country's highlands and covers 26% of annual cereal production for supplying nutritional benefits [3]. The country's Productivity of Wheat ranked first in area coverage (23.78%) and

second in total production (3.53%) in the Oromiya State. The Arsi and Bale highlands are the major Wheat producing zones of Ethiopia and are considered to be the Wheat belts of East Africa [4]. The regional average yield of Wheat is 3.52 tons per hectare [3]. Even though 85 percent of the country's Population lives in the rural areas, the performance of the agricultural sector in Ethiopia remained the weakest and it is heavily influenced by weather condition [5]. Recently as the report of [6], there was an outbreak of the Wheat disease in the critical Wheat-growing regions of Ethiopia, intimidating the crop production gains in parts of the country, following a long period of El-Nino-induced drought. Accordingly, the Rusts first spotted in Oromiya and Southern Nations, Nationalities and People's (SNNP) regions and spread to Amhara and Tigray regional states as crop-killing fungus.

In Ethiopia yield lost due to rust is reported to be 60-100%. From this Arsi and Bale regions are favorable for Stem rust development which is most dominant in the area [7]. Kulumsa Agricultural Researcher Center (KARC) which operates under the Ethiopia Agricultural Research Institution (EARI) is one of the institutions undertaking research to tackle the problem of Wheat and other crop diseases in Arsi and Bale agriculture area. In any agricultural production system, gathering and integration of related knowledge and information from many diverse sources play a key role. In practice, there is no substitute for knowledge and experience in identifying problems and choosing the most appropriate management technique for addressing them. The advice of qualified agronomist is vital in identifying threats to the crops and the most appropriate strategy for their control [8].

The problems in tackling the Wheat crop disease are inadequate knowledge about elements influencing crop rust population changing aspects. To understand these agricultural experts and researchers collected rust surveillance data and related agricultural operations regarding crops, farming activities, weather parameters, soil, and details on climatic, incidence, productivity and agricultural practices. These collected data set as a repository of information in the database. Similarly, today's development in computer technology, availability, and storage leads to the generation of a large amount of agricultural data at each agricultural sector. These resulted in the need for automatic analyzing of correlation among the attributes for acquiring and storing knowledge to support these researchers and development agents (DA). This can be achieved by integrating data mining into the computer system with reasoning capability to acquire the knowledge from available data. The knowledge acquisition from available data is automated method in which data mining techniques used as

the core activity. Therefore data mining plays vital role in analyzing the agriculture data to acquire the knowledge and patterns in data. Data mining is the technology that enables extracting hidden and interesting knowledge (patterns, models, and relationships) in very large databases. It is the essential part of the knowledge-discovery process, which combines databases, statistics, artificial intelligence and machine learning techniques for knowledge extraction [9]. The concept can also be defined as the process of discovering exciting knowledge from large volumes of data stored either in databases, data warehouses or other information repositories. Accordingly, data mining uses a number of data stores on which mining can be performed. These include relational databases, data warehouses, transactional databases, advanced database systems (such as object-oriented databases, spatial databases, text databases etc.), flat files and the World Wide Web.

The key objective of data mining is to extract knowledge from data to support and enhance the decision-making, planning, and problem-solving process [10]. Accordingly, Data mining has two key functions of i.e. prediction and description function where the Prediction function involves finding unknown values/relationships/patterns from known values, and description function provides interpretation and explanation of a large database. Classification is useful for prediction, whereas clustering, pattern discovery and deviation detection are for description of patterns in the data. Therefore data mining helps to support decision-making process by acquiring knowledge from the large set of data. But there was the problem in applying these results and taking appropriate action based on them.

Additionally, the gap existed in extracting knowledge by employing data mining techniques and using it for action. This resulted in Mining the data and analyzing the result performed by the data experts. Thus applying the data mining results to different domain areas resulted in none usability of knowledge acquired from the data mining by none experts [11]. Since the researchers and DA use conventional data analysis which is error-prone and unable to use data mining and its result for taking appropriate action. Therefore this study aimed to make the knowledge acquired from predictive model available for experts and non-experts using Knowledge Base System s which support decision makers and simplify decision making.

Knowledge Base System is the computer systems that represents knowledge about the specific problem domain and apply this knowledge to solve specific problems [9]. The knowledge to be represented in Knowledge Base System can be acquired and extracted from different sources. Accordingly, the data mining approach for knowledge-based system development sets techniques for searching through datasets, looking for hidden trends and

correlation among data (variables or attributes) which cannot be accessed by conventional data analysis. Data Mining is base for building effective Knowledge-Based System, as it uses machine learning and data statistical analysis to develop better business decisions that could be made using conventional methods. Data mining improves decision making by giving insight into what is happening today and by helping predict what will happen tomorrow. Therefore many data mining tools on the market today became powerful and help to build powerful Knowledge-Based Systems [12].

The researchers and agricultural experts are using conventional statistical models to analyze agricultural data. As clearly stated in [9], data mining technology often can improve better upon traditional statistical approaches for solving any business problems. This can be by finding additional and important variables, by identifying interaction among them and detecting nonlinear relationships. The Models that predict relationships and behaviors more accurately leads to greater profits and reduced costs. The author discussion implied the application of predictive models to solve any business problem. This can be applied in agriculture where there are plenty of data but the scarcity of experts for acquiring important knowledge. On the other hand, emergence advances in information communication technology have resulted in an ever-increasing demand to use Computers for the well-organized management and dissemination of information for decision making [13]. Keeping in view the strong need for farmers to collect important and updated information for interactive, flexible and quick decision making is an important task for agriculture management.

Numbers of the research have been conducted on the application of Knowledge Base System for different agriculture sectors. However, most of these researchers concerned with advising Knowledge Base System. For example, [14, 60] developed the prototype for knowledge based system and web-based agronomy advising system respectively. But both excluded data mining technology and recommended it as it is a powerful method. Having this and gaps in literature the objective proposed research is to build integrated predictive models and Knowledge Base System to extract patterns from Wheat datasets for detecting and treating Wheat disease. The Automated disease detection and advising system can help the users to apply the data mining technology and Knowledge Base System to detect Wheat diseases. For undertaking this research the data sets and all the supportive data were collected from Kulumsa Agricultural Research Center. Since the aim of the research is the prediction rule production classifiers algorithms namely, JRip, PART, J48 and REPTree classifiers were

selected and applied through WEKA to build predictive model. The rules generated from predictive models were integrated with knowledge-based systems. Finally, the integrated predictive model and knowledge based system prototype for detecting Wheat disease was built and tested.

1.2 Motivation to Study

Knowledge and understanding of a problem are always the first steps in identifying effective solutions to take action. Despite its high productivity, demand and caloric content cereal crop production in Ethiopia is constrained by technical and socioeconomic factors [14]. Among these cereal crop diseases are the major contributing factors toward the low yield. Recently the high outbreak of Wheat diseases resulted in the loss of productivity throughout Wheat production area. This problem needs sufficient and knowledgeable experts to predict, identify the diseases and describe the methods of treatment and protection at the early stage. The objective of this research is to build integrated predictive mining model and the knowledge-based system for Wheat disease detection that assists agriculture experts and development agents in making timely decisions. To achieve this, important knowledge was acquired automatically by mining by analyzing Wheat dataset. Additionally, the proposed research conducted interview and document analysis to find additional and detailed knowledge from experts**.** This helps to acquire knowledge about Wheat diseases, the symptoms and treatment methods undertaken to control it. As seen from the current practices in the agriculture, there is no automated predictive model and advising Knowledge Base System, which will help the researchers and development agents to show those Wheat crop diseases have a high degree of prevalence, severity and treatment activities. Therefore, the aim of this research is to identify patterns which are describing Wheat diseases and represent these patterns in Knowledge Base System and to make the data mining findings easily accessible to the experts and non-expert users.

1.3 Statement of the Problem

The underlying research problem that necessitated this research is the outbreak and high prevalence of **Wheat diseases** in critical Wheat-growing regions of Ethiopia. It is threatening recent crop production gains in parts of the country, following a long period of El-Nino-induced drought. A late survey taken by agricultural research teams shows the rust has been spotted in Oromiya and Southern Nations, Nationalities and People's (SNNP) regions [6].

Then, this crop-killing fungus was spread to Amhara and Tigray regional states. This resulted in the need for efforts of all agricultural stakeholders, government strategy, experts, farmers, and researchers. From the evidence of [15], in Ethiopia, agriculture experts may not be available, and accessible always to every farmer. This is mainly due to a scarcity of agriculture extension worker the so-called development agents. Accordingly, each development agent has to attend on average 1090 farmers, which was beneath the expected standard of agriculture transformation in the country. Further, it also stated that there are only 14 percent research experts of research institutions are focusing on controlling pests and crop diseases. The Agricultural production system has evolved into a complex business requiring the accumulation and integration of knowledge and information from many diverse sources.

For agricultural sectors to persist competitively, the modern farmers often rely on agriculture experts and advisors who provide information for decision making. Unfortunately, agricultural experts' support is not always available when the farmer needs it. To solve this problem, [18] recommended knowledge based system as the powerful tool for improving and supporting decision making in agriculture. Knowledge Base Systems are a subdivision of artificial intelligence in which computer program attempts to replicate the reasoning process of human expert [17].

Additionally, the author stated applications of the expert system in agriculture mainly focus on disease identification and pest controls. This indicated the Knowledge Base System can be used as a tool to disseminate available best practices to overcome the effects of diseases. The Significance of Knowledge Base System technology has been recognized and realized in the field of agriculture. Numbers of the research have been conducted on applying the Knowledge Base System for different agriculture sectors excluding data mining technology. Abdulkarim [18] Recommended local researchers to diversify knowledge acquisition techniques to automatic for developing knowledge based systems rather than conventional methods in many fields of study. Additionally, the study implied the application of this recommendation in other areas where there is the shortage of domain experts to acquire knowledge but there exist tremendous data in such way knowledge can be acquired automatically from the data. Furthermore, [19] also used integrated knowledge acquisition technique which means both manual and automated to construct the Knowledge Base System and recommends the method to be extended in another domain areas. Agriculture research institution like Kulumsa contains a huge repository of data for generating the knowledge. As

to this research, no attempt has been made to mine their data for extracting hidden knowledge and patterns in these data.

The researchers explored the application of Knowledge Base Systems and Data Mining in agriculture for disease detection and treatment alone. From the recommendation of [14] and limitation of knowledge based systems developed by the manual elicitation of the knowledge and criticism on Data Mining results by [18]; it is understandable that integrating mining predictive model and Knowledge Base System is essential to deploy knowledge extracted from mining model for developing the knowledge based system. The Sufficient knowledge must be acquired about the problem to be solved. This Knowledge can be acquired from diverse sources such as interviewing the domain experts, analyzing manuals and related documents, observation and other methods [20]. Subsequently, tacit knowledge is personal knowledge and expert may not express all the knowledge he/she knows during the interview, there is hidden knowledge about the problem. To alleviate this problem, automatic knowledge acquisition is proposed by [18, 21]. Data mining has been proposed for extracting hidden and previously unknown knowledge from datasets by different researchers [18].

The above mentioned and many researchers attempted to apply knowledge based systems in the area of agriculture to support agricultural experts. However, as to the knowledge of proposed research, no study has been attempted to apply an integrated predictive model and knowledge based system for detection and treatment of Wheat diseases caused by different pathogens. Generally, all the problems stated above have encouraged the researcher to build a predictive model and the knowledge based system which assisting both researchers and agricultural experts or development agents (DA) decision making in detecting and treating Wheat diseases. This helps to make the quality decision on daily needs of farmers. Therefore, the aim of this research is to build and apply integrated Predictive Model Mining and Knowledge Based System for Wheat disease detecting and treatment. To the end, the proposed research will answer the following research questions

RQ1: Which data mining technique is the most appropriate to build the predictive model?

RQ2: What are the patterns that characterize Wheat crop diseases?

1.4. Objectives of the Study

The Proposed research contains both general and specific objectives to be achieved.

General Objective

The general objective of the study is to investigate the construction of a predictive model to extract hidden knowledge and integrate with Knowledge Base System for detecting and treating a Wheat disease.

Specific Objectives

To accomplish the above stated general objectives, the following specific objectives are developed:

- To review the literature and analyze state-of-art in data mining, knowledge base system, and previously solved problems in both.
- To explore and identify suitable Data Mining techniques and build predictive model on Wheat data set
- To acquire knowledge from predictive model and domain experts that will be used for Knowledge Base System construction.
- To build integrated prototype Knowledge Base System from predictive models
- To explore applicability of integrated predictive models and KBS in agriculture.
- To test and evaluate performance and user acceptance of developed prototype.

1.5. Scope and Limitations of the Study

The scope of this research is limited to focus on the Agricultural sector of Wheat disease. The proposed research aimed to build the predictive model and integrate with the knowledge based system to support detection and treatment the Wheat diseases. Additionally, the research acquired expertise knowledge from domain experts and analyzed documents and manuals to have adequate knowledge of domain area. Furthermore, the research work developed integrated mining predictive model and knowledge based system prototype for supporting detection and treatment the Wheat diseases. There are many constraints to crop productivity; the seasonal outbreak of Rusts becomes the common problem which is the initiative for this research. The research focused at Kulumsa Agricultural Research Center which is conducting research on crop and farming related activities at Arsi zone areas of production. Therefore, the study will not include other types of crops and the study used the data sets of Wheat disease. On the other hand, the data collected at research center was not in

a suitable format for this study, collected data was not complete and well organized, since different researchers stored data on their own computer, problem of collecting data from the center and the collected data set does not include all types crop diseases, are some limitations to the proposed research work results**.** The research also limited to experiment with alternative data mining tools other than WEKA and set its comparison with other tools. Lack of sufficient literature and related works especially in the application of data mining in crop diseases and lack of prior experienced and skilled professionals in data and data mining application at agriculture sector to provide proper and timely feedback are some limitations of the study.

1.6. Significance of the Study

The main outcome of proposed research result will be enhanced decision making to mitigate and control problem of Wheat diseases timely. The findings of this proposed research will benefit the agriculture researchers and development agents in such a way they can easily get disease detection and diagnosis advice to manage Wheat diseases from the integrated Knowledge Base System which stores the facts and rules used to solve problems. By proper use of the knowledge acquired from predictive model and domain experts, the knowledge-based systems will increase productivity, and enhance problem-solving ability in a flexible method. This can also document knowledge for future use and training which may lead to having improved problem-solving ability. On the other hand, it will be beneficial to farmers indirectly as it can be used for consultancy and training tool for researchers and development agents. Finally, it will reduce the existing knowledge gap observed in remote areas where skilled agricultural professionals are scarce and may not be available. It also serves as being a knowledge sharing and training tool for new recruited development agents (DA) and researchers. This can improve the skills of agriculture researchers and development agent workers in Wheat disease identification and decision making on its treatment.

Benefits and Beneficiaries of the study

These research works will give benefits to stakeholders in the agriculture sectors and researcher in the following ways:

- ✓ To help the experts and non-expert in Wheat disease identification, diagnosis, and treatment. They also improve their knowledge and skills to be effective in identification and treatment of Wheat diseases as experts' knowledge is readily available for them.

- ✓ To help agricultural researchers and development agents in analyzing the relationship between attributes or variables to easily understand factors contributing to Wheat Diseases. This leads to the enhanced decision which may help farmers indirectly as they are dependent on advice and information for their daily activities on production.
- ✓ To apply the knowledge acquired from integrated data mining models and knowledge based system in agriculture especially in areas with a shortage of domain experts or development agents.
- ✓ The researcher also benefited by gaining experienced knowledge of both buildings the predictive mining models, Knowledge Base System, and planning of the research.

1.7. Organization of the Thesis

This section dealt with the organization of thesis as the following. Chapter one considers an introduction to the study by giving highlight about the area to which the study is focused. It also describes Wheat crop and its diseases, Data mining and knowledge based system. Finally, the section also discusses the statement of the Problem, objective the study, significance of the study, scope, limitation and significance of the study.

Chapter two is mainly dedicated to discuss the literature review. In this section, a detailed discussion about Data mining and its tasks in agriculture were provided. The general concept of Wheat disease, its detection and treatment merely discussed. Further Discussion about Knowledge Base System, reasoning mechanisms, knowledge acquisition and representation in the knowledge base, also included.

Finally, the works of other scholars on data mining, Knowledge Base System and integrating both were reviewed and presented as related works. Chapter three dedicated to present the research methodology used for conducting this research. In this section, the research area, period tools and the design set to conduct the research was discussed in detailed form. Additionally, the data collection and knowledge discovery techniques used in this study has been highlighted clearly. Chapter four stated details of experiments and analysis of the results obtained. It also deals with disease detection, data understanding, processing steps, and experiments in building predictive models and its analysis of results was discussed in details. The results obtained from classifier algorithms are analyzed, interpreted, evaluated, compared with each other and finalized. Further discussion of the integration of dataset induced rules from data mining techniques with Knowledge Base System also included in this section.

Chapter five is all about implementation results from predictive models to construct Rule-based Wheat disease Detection and treatment Knowledge Based system prototype. Mainly the basic functionality of the Knowledge Base System prototype, performance evaluation, and user acceptance testing were included.

Finally, Chapter six is dedicated to present the final conclusions and recommendations forwarded by this research work. In this chapter based on the result obtained from the study, the researcher‘s concluding remarks and recommendations for future works also presented.

CHAPTER TWO

LITERATURE REVIEW

This research work has reviewed many related literatures (books, journal articles, Conference proceeding papers, and the Internet) in order to have a detailed understanding, support background information and identifies the research gaps in the literature on the present researches.

2.1. **Wheat Crops in** Ethiopia

Ethiopia has favorable Wheat growing climatic environment which makes the country the largest Wheat producer in Sub-Saharan countries. Wheat is primarily produced by subsistence farmers under rain-fed conditions and the 4th important cereal crop after maize, teff, and sorghum in the country. Although Ethiopia is the largest Wheat producer in sub-Saharan and reliant on foreign Wheat imports to satisfy country's annual internal demand [22]. Therefore, the Ethiopian government through the Agricultural Growth Program is active in efforts to improve the production and productivity of Wheat to increase domestic supply in the country. In Ethiopia Wheat is primarily grown in the Oromiya, Amhara, Tigray and Southern Nations, Nationalities and Peoples (SNNP) regions. These regions account for more than 90% of national Wheat production, of which 50% is the contribution of Oromiya state. In 2012/13 the estimation showed under cultivated area the Wheat was more than 1.5 million hectares of land with annual production of nearly 3 million tons.

It is estimated that over 4.5 million farmers are annually involved in Wheat production in the country. Thus, 65% of the Wheat produced is of the bread Wheat type, while 35% is of durum Wheat type and is homegrown to the country [22]. Besides its high productivity and need the Wheat production is intimidated by recent outbreaks of rusts throughout the regions of the country. Mainly Arsi and Bale's regions are favorable for the development of Stem rust dominantly invading its production [7]. According to Ethiopia statistical agency of 2011/2012, the average Wheat production capacity has grown to show some improvements changing between 1.84 to 2.03 tons per hectare on average, which is still extremely beneath international standard levels. These are due to constraints of Wheat production like diseases and pests particularly rusts. Many of widely cultivated commercial Wheat varieties such as Kubsa and Galema are susceptible to rusts. In an effort to increase the productivity of Wheat,

the constraints to the Wheat production are categorized into technical and socioeconomic factors. The Technical constraint includes Diseases, low soil fertility, weeds, and varieties. As well as, the socioeconomic constraints are Unavailability of improved inputs, Seasonal labor shortages, Draft power shortage, Land shortages, Low prices and lack of credit. Wheat diseases like Stem rust and leaf rust are principal disease problems to durum Wheat while Stripe rust/Yellow rust and Septoria tritici are the main problems with the bread Wheat. These rusts are mainly reported in kulumsa and Holleta agriculture areas first time. Similarly, major diseases and pest of the Wheat crop in Ethiopia can be categorized into fungal, bacterial, viral, nematodes and physiologic and genetic disorders [23]. The details of these diseases have been elaborated and attached to Appendix I of the paper.

2.2. Data Mining and Knowledge Base Systems

The growing of the computers, computing technologies, and internet emerged in the generation of a large amount of data at manufacturing institutions, government agencies, scientific institutions, and business have an enormous resource to collect and store data. But in the practical small amount of data will be analyzed and used for solving business problems. These are due to the reason that many of institutions, companies, and agencies used conventional data analysis and faced incapability of managing and using large records in their decision. Data mining sets solution for the business problems using historical data. Many scholars' defined data mining from different perspectives, for example, [24] defined data mining as an extension of traditional data analysis and statistical approaches incorporates analytical techniques drawn from a variety of disciplines including pattern matching and areas of artificial intelligence such as machine learning, neural networks, and genetic algorithms. The author also stated data mining is not limited to numerical analysis only. While many data mining tasks follow a traditional and hypothesis-driven data analysis approaches. It is Commonplace to employ an opportunistic, data-driven approach that encourages the pattern detection algorithms to find useful trends, patterns, and relationships. Similarly, it also defined as the process of extracting useful knowledge from a large amount of data. Many people may know data mining as a synonym with the Knowledge Discovery in Database (KDD).

But KDD is a general term to refer overall process of discovering useful knowledge from data, where data mining is a particular step [10]. The authors of [24] settled the idea that data mining is the essential step in knowledge discovery in the database and approved more

general term data mining as the process of discovering interesting knowledge from large data sets stored in databases, warehouses, and other depositories of data. Data mining is the practice of exploring data from different views and summarizing it into useful information that can be used to increase revenue and reduce costs, and/or for both. Data mining software is one of the analytical tools for analyzing data and allows users to analyze data from different ways, to classify, cluster, visualize, identify and summarize the relationships in data. It can also be the process of finding correlations or patterns among dozens of fields in large relational databases.

The method allows the analyzer to acquire, previously unknown or hidden knowledge about relations and patterns in the behavior of the investigated object from data [25] Recently Data mining has shown the dramatic increase in the amount of information and data stored in different electronic format. The issue is how to maintain these data and discover knowledge used for helping the decision-making process. It has attracted a great deal of attention in recent decades because of the extensive availability of large amounts of data and the imminent need for turning such data into useful information and knowledge. These acquired information and knowledge can be used for applications ranging from market analysis, fraud detection, disease detection, intrusion detection and customer retention, to production control and science and exploration in any fields of study [10].

2.2.1. Objectives of Data Mining

[26] "Described the ultimate objectives of Data mining as to identify valid, interesting and potentially useful and understandable correlations and patterns existing in data. Finding the useful pattern in data is known as knowledge extraction, information discovery, information harvesting, data archeology and pattern processing". Data mining was predominantly used by statisticians, database researchers, and the Management Information Systems and business societies. As early stated above term KDD is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is taken as a particular step in the process. The additional steps in the KDD such as data preparation, data selection, data cleaning, and proper interpretation of the results are to ensure useful knowledge is derived from the data.

2.2.2. Data Mining and Knowledge Discovery Process Models

The accessibility and abundance of today's modern computers and internet resulted in the availability of large quantity of data. These made Knowledge Discovery and Data Mining a matter of considerable importance and necessity in any field of study. Thus knowledge discovery process models define sequential steps with eventual feedback sets of loops that should be followed to discover knowledge. Knowledge Discovery in Database /KDD is defined as the process of identifying valid, novel, useful, and understandable patterns from large datasets. Consequently, Data Mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and algorithms to explore data and discover significant patterns in data, which are the core of useful knowledge. These mathematical models are used for understanding phenomena from the data, analysis, and prediction. These models support data practitioners and institutions by providing a roadmap to be followed in the planning and using data mining projects [27]. There are many KDD process models. Fayyad and hybrid KDD process model have been preferred to be discussed in this study. The hybrid KDD process model is interactive and iterative processes that can move back to adjust previous processes at each step. These process models can be initiated at understanding application domain and identifying the goal of KDD from the customer's perspective. The process models developed by Fayyad et al stated that the Knowledge Discovery is non-trivial processes of knowledge exploration which consist six steps [28].

These includes understand and selecting target data which involves selecting dataset or focusing on data samples or subset variables needed to be discovered. The next step is Data cleansing and preprocessing which involves removing redundancies and noise, and handling missing data values. Then, data transforming is preceded where data is transformed into a suitable format for mining. This can be summarizing or aggregating data which leads to data mining phase. Data mining involves searching for patterns of interest by building a model and evaluating or interpretation and using the results obtained from the models. This includes tasks of data classification, decision trees, regression, and clustering. The fifth step entails visualization of extracted patterns which involves interpretations of the results obtained from the model. The final step is knowledge representation which involves using obtained knowledge directly, incorporating into another system for further actions or documenting and reporting to the interested body. Figure 2.1 illustrates the KDD process steps.

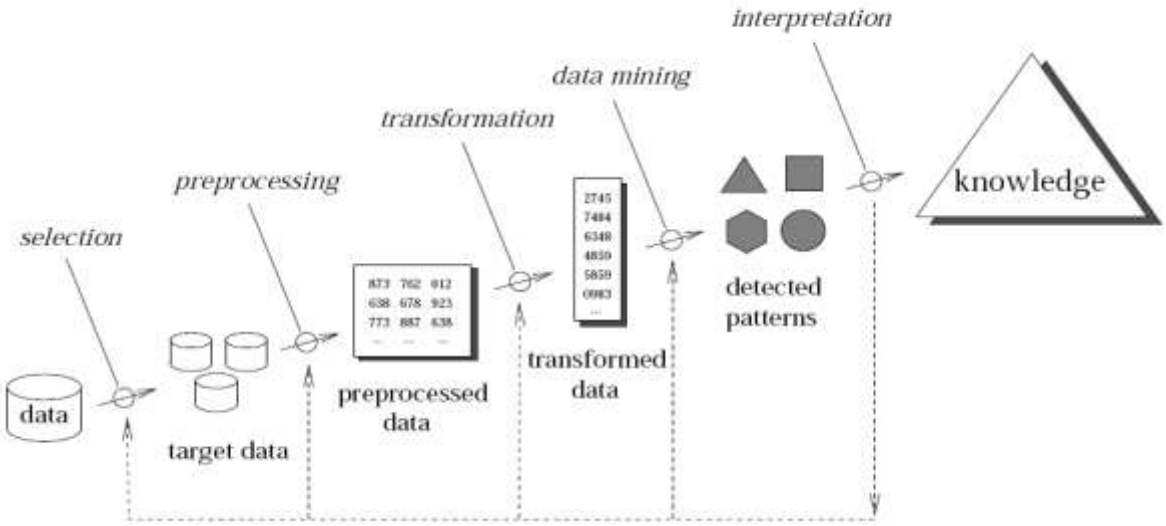

Figure 2-1 Knowledge discovery process models adopted [29]

Moreover, Fayyad [28] also defined data mining as one step in the KDD process consisting of applying data analysis and knowledge discovery algorithms under acceptable computational efficiency and limitations to produce a particular enumeration of patterns over the data. Accordingly, Data mining is the step that is concerned with the actual extraction of knowledge from data. To emphasize the necessity that data mining algorithms need to process large amounts of data, the desired patterns have to be found under acceptable computational efficiency. Data mining can also be defined from a business perspective as the process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions [30]. Therefore, Data mining is the main part of knowledge discovery process which helps to extract knowledge from data automatically. The knowledge and information obtained from Data mining can use in different applications. Another KDD process model is hybrid process model which is originated for both academic and industrial projects. It combines both aspects and research-oriented model. The hybrid process model consists of six steps to be followed in discovering knowledge from data. These include: Initially begins with

Understanding the business or problem domain: this is the initial step which involves working closely with domain experts to define and understand the problems. It also determines the project goals, and learns current solution and translating business project goals into Data mining goals.

Understanding the available data: it involves collecting sample data, deciding the data size and format to be used for analyzing with data mining. The domain experts' knowledge needed to check for completeness, redundancy, missing values, the plausibility of attributes values and verification of usefulness of data with the Data mining goals to be attained.

Data preprocessing and preparation: this involves deciding data be used by sampling, running correlation and significance tests, cleaning, correcting for noise and missing values. Further, it can be processed for features selection and extraction algorithms.

Data mining: the data mining step involves the miners used DM method to derive knowledge from data. It includes searching for patterns of interest and relationship among the data attributes by building a model and evaluating or interpretation and using the results gained from the models. This includes classification, decision trees, regression, and clustering.

Evaluation of discovered knowledge: this step includes understanding results, checking whether it is novel, interesting and interpretable by domain experts.

Use of the discovered knowledge: includes planning how and where to use discovered knowledge. The discovered knowledge from the data mining results will be used for decision making purpose [29]. Since this research was conducted for academic purpose and hybrid process model is research-oriented it has been adopted and used as a research method to complete this research.

2.2.3. Basic Data Mining Tasks

The functionalities of Data mining are used to specify the kind of patterns to be found in data mining tasks. The data mining tasks can be classified into predictive and descriptive categories [24]. Similarly, [31] stated that a Predictive model makes estimation about values of data using known results found from different data: while the Descriptive model identifies patterns or relationships in data with the objective of identifying the natural grouping of data. Unlike the predictive model, a descriptive model serves as a way to discover the properties of the data examined, not to predict new properties. Further, prediction models are supervised data mining task, while descriptive models are unsupervised and visualization aspect of data mining. Additionally, most of the data mining techniques are based on inductive learning, where a model is constructed explicitly or implicitly by generalizing from a sufficient number

of training sets. The underlying statement of the inductive method is that the trained model is applicable to future cases. Some predictive models also provide an understanding of data [27]. For this study predictive models used to analysis agriculture data and acquire knowledge. Some basic DM tasks have been discussed as the following.

➢ Classification

Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict classes of the object whose class label not known. It determines the class of an object based on attributes or variables available in the data. Thus it finds out the relationship between predictor value and the target value of data. The model built based on the analysis of set training data set or data objects whose class label is known [10]. The predictive mining tasks infer from current data to make a prediction of relationships and patterns among different variables and parameters in data sets. Classification is commonly used data mining technique for prediction. In classification technique, the historical data set can be split into the training set for model building and testing set for testing the built model. Some application of classification techniques are stated [32] as the following points:

- Diagnosing particular diseases
- Detecting fraudulent activities.
- Identifying financial or personal behavior.
- Determining whether a particular credit card transaction is fraudulent and
- Assessing whether a mortgage application is a good or bad credit risk and similar activities can be undertaken by classification.

The most commonly used classification techniques are Artificial Neural Networks, K-nearest neighbor, the Naïve Bayes technique, Decision trees, Support Vector Machines and Rule-based learning and Production rules are included in the classification.

➢ Decision Trees

Decision trees are predictive models that can be used to represent classifiers. In data mining, decision trees represent both classifier models and regression models [27]. Similarly, Decision tree is typical learning methods that construct trees from the set of input and output samples of data. These learning systems of decision tree adopt a top-down strategy to search

for a solution for a given problems [33]. The Decision trees induction learning algorithms Commonly used in machine learning and applied statistics are ID3, C4.5, J48, and Classification and Regression Trees (CART) are used to construct decision tree in top-down recursive divided and conquer manner and most of the decision tree induction algorithms follow divided and conquer approach [10]. The decision tree starts training set of instances and their associated class labels. These training sets recursively separated into subsets as the tree built. The key aim of the decision tree is to develop classification rules from the data in training dataset [34].

This can be done by the process called splitting the values of attributes into testing attribute value and creating branch tree for each attribute values. For continuous-valued attributes test is normal whether the value is less than or equal to or greater than given value to the split value. The authors of [10], [34], and [35] were sorted decision tree constructing algorithms as ID3, C4.5 and CART in their respective order of invention and usage. The attractiveness of decision trees is due to the fact that, unlike neural networks, decision trees used to represent rules. These Rules can easily be expressed and understood by humans or even directly used in simple database access languages. In addition, some properties of the decision tree are listed by [35]. Decision trees:

- Follows Divide-and-Conquer strategy, which has weaknesses of fracturing and diminishing training data and Learns discriminating rules,
- Follows general-to-specific search and Noise-tolerant and
- Learns with positive and negative examples.

Larose et al [30] set the requirements to be met for employing decision tree algorithms. First, decision tree requires predetermined target variables and training data set to give algorithms with target variables value. Second, the training set to be used should be rich enough and varied, providing the algorithm with a healthy cross-section of types of record to be classified in the future. As the decision trees learn by example for training sets lacking definable subset of records, classification and prediction will be difficult. Thirdly, the target attribute classes must be discretely valued which are clearly defined as either belonging to a specific class or not. Decision tree classifiers have good accuracy depending on the data set at hand.

Decision Tree classifiers construction does not need any domain knowledge or parameter. This makes it appropriate for exploratory knowledge discovery and acquired knowledge is

intuitive and generally easy to adapt by humans, as the learning and classification steps of decision tree induction are simpler and fast [10]. Decision tree induction algorithms for classification are applied in the area of medical diagnosis, manufacturing and production, financial analysis, astronomy, and molecular biology [33].

J48 Classifier: J48 is one the decision tree building classifier which is an employment of Quilan algorithm or C4.5 for classification of a given datasets. In J48 classifier's nodes are to represent discrimination rules acting on selective features by the iterative splitting of data using depth-first strategy [33]. Further, the algorithm used each attribute of the data set to make a decision by splitting the data into smaller subsets. All the possible tests are considered during decision making based on information gain value of each attribute in the instance of data set.

➢ Rule-Based Classification

Rule-Based Classification is separate-and-conquer methods which are iteratively processing data to generate the first rule that covers a subset of training examples and then removing all covered examples by rule from the training sets. The process repeated until no examples left to be covered in training set. Rule-based classifiers group instances by using a set of "IF-THEN" format to represent discovered rules [36]. This can be written in the following

Rule :(Condition) -> X; Where Condition is a conjunction of attributes and **X** is called a class label. The rule comprises antecedent or condition as left-hand side and consequent/action as right-hand side. A given rule covers instances if the attributes of the instances satisfy the condition of the rule. The basic unit and format of knowledge in rule-based reasoning is the rule. These rules are reasonably natural for humans and are easily understandable for both programmers and domain experts. But, the accurate description of the domain expert's knowledge in simple rules is often difficult. PART and JRIP algorithms are some of rule-based classifiers for building the classifier models [36].

PART: PART is rule-based classifier algorithm to produce sets of rules called decision lists or partial decision trees which are ordered set of rules. The algorithm match rules by comparing new data to the rules in the list and assign item category of first matching rule. It built a partial C4.5 decision tree in iteration and adds the best leaf to the rule. The algorithm is a combines C4.5 and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) rule learning and use the separate-and-conquer method to identify the rules [37].

JRip: JRip is another rule-based learning algorithm. Its rules are generated and pruned for each and every class available in the training set of the data. The acquired knowledge represented in the format of IF-THEN production rule. The IF-THEN rules are high level and symbolic knowledge representation contributing to the clarity of the discovered knowledge. JRip classification is based on the construction of a rule set for all positive instances covered by partitioning the current set of training instances into two subsets namely a growing set or rule growing and a pruning set or rule pruning. 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 negative instances are covered. Following this, the algorithm substitute or revises individual rules by using reduced error pruning to increase the accuracy of rules. To prune a rule the algorithm takes into account only a final sequence of conditions from the rule and sorts the deletion that maximizes the function [38]. In general, JRip is a propositional rule learner and follows the repeated Incremental Pruning to Produce Error Reduction (RIPPER).

2.2.4. Application Data Mining Techniques in Agriculture

The most significant achievements of research in information technology resulted in the development of techniques allowing to model information at a higher abstracted level. Data mining makes information technology to develop and utilize data for making a decision by extracting regulations, patterns, and models from large databases. This resulted in the broadly improved application of Data mining which attracts the attention of many scholars to apply data mining in agriculture. Agriculture yield is affected by different factors which can be analyzed from historical data by statistical and computational methodologies. Applying such methodologies and techniques to historical yield data leads to obtaining information and knowledge which is helpful to farmers and government organizations for making better decisions and policies. These directly resulted in helping agriculture to increase its production [39]. Data mining techniques can also be applied for estimating yield. This help farmers and farming institution to have awareness on which crop to on depend based their productivity and factors affecting them [40]. In general, the above discussed and many scholars applied data mining techniques and proven its application in agriculture. Consequently, Data mining can be applied in agriculture for the following;

Crop Yield Estimation: - an accurate estimation of crop production and risk management helps agricultural companies in planning supply chain decision like production scheduling.

Business such as seed, fertilizer, agrochemical and agricultural machinery industries also plan production and marketing activities based on crop production estimates.

These can be helpful in two ways both for farmers and government in decision making namely; helps farmers in providing the historical crop yield records with a forecast reducing the risk management and government also used in making crop insurance policies and policies for supply chain operation [41]. Estimation of damage caused by the pest, data mining for meteorological data analysis, mushroom grading, and spatial data mining to reveal interesting patterns related to agriculture are some applications of data mining.

Customer Behavior Prediction: agricultural enterprises analyze their data from different perspectives and find the relationships and connection among them. Since agriculture enterprises collect very plenty and diverse data, organization and integration of these data enable creation of agricultural information systems**.** This provides agriculture opportunities for exploring hidden patterns and relationships among these collections of data. These, in turn, used to determine the condition of their customers in markets and support decision making. Therefore application information technology through data mining increasingly provided systematic assistance in solving agriculture problems [42]. Consequently, the author claimed that application of data mining in agriculture supports agricultural enterprises to use their data for the following purposes.

- To identify valuable clients and potential customers in the future,
- To divide the markets and customers into segments,
- To investigate causes of customer behavior change and successful tactics to acquire and keep customers,
- To define prices for different individual customers segments and identify poor payers,
- To create customers profiles that the enterprise desire to keep and acquire customers.

2.3. Knowledge Based Systems

The Knowledge Base Systems are the important family member of an artificial intelligence systems group. The terms Knowledge Base System, expert systems, and decision support system are similar and can be used interchangeably [43]. The Knowledge Base System is a sophisticated computer program with highly expertise knowledge and acts as an expert on request without more wasting time, at any time and anywhere. The KBS utilizes knowledge properly, to increase productivity by replacing the scarce expertise and by enhancing

problem-solving abilities in a flexible manner. It also documents and stores the knowledge for future use and training. This leads to enhanced decision making and problem-solving process in any kind of business area. The KBS is computer system comprises of a collection of expert knowledge from domain experts with utilities of facilitating the knowledge retrieval in response to specific queries, along with learning and justification to transfer skill from one domain of knowledge to another. Mainly, it focuses on using knowledge-based techniques to support human decision making, learning, and action constituting the capability of cooperating with human experts. It can be used for problem solving, training, assisting users and domain experts for whom it may be applied.

Unlike traditional computer systems such as, transaction processing and management information system which do not understand data and information, the KBS generates and uses knowledge from data, information and knowledge by understanding data and information being processed to make a decision [44]. On the other hand, knowledge sources can be documented and undocumented source, which obtained from individuals. This kind of knowledge identified and collected by human senses or machines such as sensors, scanners, cameras, pattern matches intelligent agents. These variations in sources and types of knowledge resulted in the complexity of the knowledge acquisition**.**

2.3.1. Advantages of Knowledge Base System

The application Knowledge Base System has gained understandings and recognition in different domain areas. It is advantageous over many traditional computer-based information processing systems. Many authors specified advantages of KBS from different perspectives. Thus some advantages listed in [45] are selected and discussed as the following**.**

KBS achieves self-learning and ease of updates: it updates its knowledge from experience with the help of systems' inference engine. This can be either use automatic machine learning or manual editing of the knowledge by knowledge engineer or editor in the knowledge base.

Provide low-priced solution and easy availability of knowledge: KBS development can costs once for building system and copied into different places of application to make the use of knowledge easy. These realize the improvements in the dominance of the knowledge experts in domain area and simplify knowledge acquiring and usage, which resulted in reduced cost, time, human expertise and error.

Provide permanent documentation of knowledge: the knowledge in KBS can be elicited and collected from domain experts and relevant documents. Further, it can be represented and

transferred into the knowledge base to solve certain problems. This makes knowledge to be stored and used for long-term from documented and undocumented knowledge.

Provide justification for better understanding, consistency, and reliability: its' reasoning and justification support the reliability of domain expert on decision making. Well, understanding and decision made raise quality and reliability of the system. This resulted in increased in level and amount of knowledge which leads to the right decision and reduced error with appropriate justification in the decision.

Effectiveness and efficiency: Since knowledge-based systems are computer-based systems, they have efficiency-directed elements such as speed, accuracy, control, and permanent content storage.

2.3.2. Limitation of Knowledge Base System

As discussed in the above Knowledge Base System has many advantages over the classical computer-based information processing systems. However, it also limited in following ways [45].

Partial learning and development method: to make the decision KBS can explain and learns from experience through updating its knowledge. But, the represented knowledge may or may not be known fully which leads to partially learning of KBS. Similarly, for information system developments there is commonly accepted method, common guidelines and lifecycles are set for developing all kinds of computer-based information systems. But, no common development models set to develop Knowledge Base Systems.

Weak support of methods and heuristics: Knowledge Base System cannot respond if the problem is out the systems knowledge. This resulted in non-response to problems do not represent in KB and does not support the solution. Thus the knowledge engineers are responsible to develop heuristics to solve this limitation.

Knowledge acquisition, Creativity, and innovation: the knowledge acquisition is transfers of knowledge from the source into KBS in a suitable format and representing into the computer in an understandable form. From its nature knowledge is personal and extracting knowledge embedded in human mind can be very challenging activity in Knowledge Base System development.

Testing, certifying and standards for Knowledge Base System: as the Knowledge Base Systems operated in specific problem domain. It also limited in absence of standardization in its acceptance.

2.3.3. Architecture of Knowledge Based System

The advances and availability in computing power and resources turn the attention of demanding tasks into the need for more intelligence. Manufacturing industries and others come to be Knowledge oriented and depend on expert's knowledge for decision making. Thus, the KBS act as an expert on demand with the capability of saving money and time, at any time, anywhere. It also considered as a productive tool allowing users to function at a high level in increasing the consistency of work [46]. Moreover, KBS composed of two main components i.e. knowledge base and inference engine or search program and other components are also available to interact. The Knowledgebase is the repository of domain expert's knowledge and meta-knowledge in different forms, including empty workspace to store temporary results and knowledge pieces; while the inference engine is a software program that infers from the knowledge available in the knowledge base. The KBS's creditability also depends on the Explanation and Reasoning of the decision made or suggested by the system as of the human expert's. Learning is also an essential part of Knowledge Base System in which knowledge can be updated either manually or automatically by machine or machine learning. Further, the Knowledge Base System should have an appropriate user interface with natural language processing facility to retrieve knowledge from the inference engines. In general, [46] divided KBS into the knowledge base, inference engine, user interface with explanation or reasoning and self-learning facilities. Although [47] also agreed with the division of KBS components containing a knowledge base and inference or reasoning engine, user interface, update and explanation facilities, user or an expert and knowledge engineer and knowledge acquisition methods. As to the author, KBS is the computer program with the capability of simulating the judgment and behavior of human with high-level knowledge and experience in the specific field of study. It is domain-specific knowledge and applies forward and/or backward reasoning, heuristics and explanation capabilities to solve specific problems. Other scholars and the author suggested components and architecture for the Knowledge Base System. Figure 2.2 illustrates the components of Knowledge Base Systems.

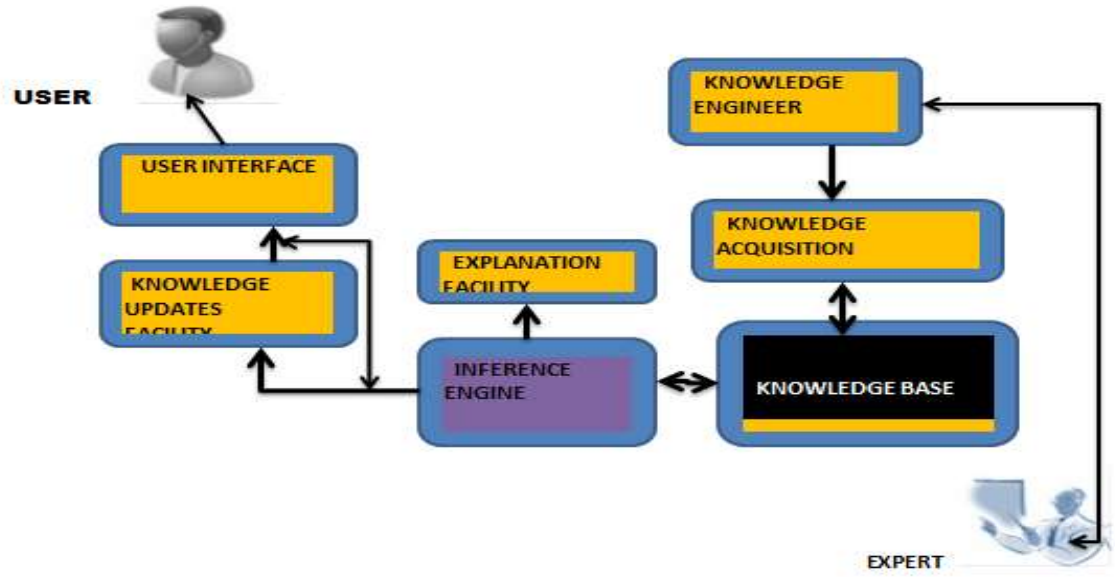

Figure 2-2 KBS Architecture Adopted [47]

2.3.4. Knowledge Engineering Process

The human expert stores knowledge in a very abstracted way in his/her mind. The knowledge engineering is an artificial intelligence that creates and applies rules to data to mimic the thought process of human experts in solving problems [48]. The process looks at the internal structure of task or decision to identify how the conclusion is reached. The knowledge engineering can also be defined as the process of transferring human experts' knowledge into a computer program in attempting to mimic the problem solving activity of human experts. Originally the knowledge engineering required to transfer the expertise of problem-solving human experts into a program that could take in the same data and come to the same conclusion.

This approach is dominated early knowledge engineering attempts and focused on creating a system that will hit upon the same results as an expert do. Further, the Knowledge engineering is the process to obtain experts' knowledge to build Knowledge Base System. It involves all the technical, scientific and social aspects involved in building, maintaining and using Knowledge Base System [49]. Knowledge Engineering can also be defined from two perspectives i.e. narrow and broad. The Knowledge Engineering can be dealt with knowledge acquisition, representation, validation, inference, explanation, and maintenance from the narrow perspective. In the contrary, the KE from the broad perspective defines the entire process of developing and maintaining intelligent systems. The major goal of Knowledge Engineering is

to help experts be clear about what they know and document the knowledge in a reusable format. The knowledge from a human expert is base for the development of Knowledge Base System [50]. Similarly, the development of KBS is the integration and involvement of many components including knowledge engineer who acquires, transfer, and represents domain experts' knowledge in computer understandable format, users, and experts, interface for users group and processes [46]. Figure 2.3 depicts the components and Development of a KBS.

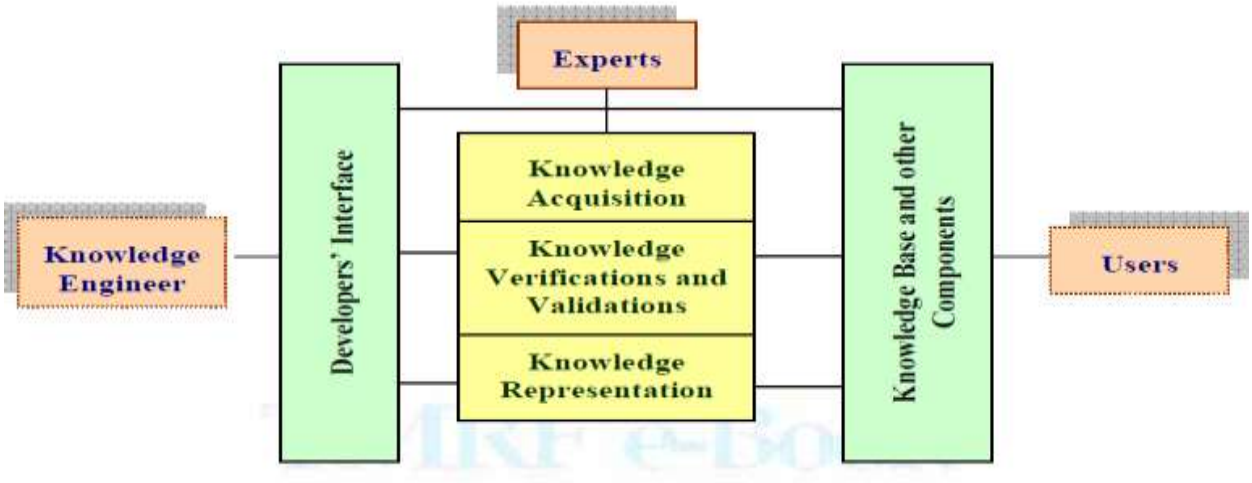

Figure 2-3 Development of Knowledge-Based System, adopted [46].

- **Knowledge Acquisition**

Knowledge acquisition is the process by which the knowledge engineers acquire knowledge from domain experts, group of experts, documentation or textbooks, manuals, multimedia, databases, reports, and from the web. This acquired knowledge used to build complete, accurate and well-organized knowledge based system and can be specific to the problem domain. In general, Knowledge can be classified in many different ways. The knowledge can categorized as tacit, explicit, factual, procedural, commonsense, domain knowledge and Metaknowledge [46].

Tacit knowledge: is the knowledge stored in the subconscious mind of human experts and cannot be easy to document.

It is highly personal and hard to formalize, and hence difficult to represent formally in system. Subjective insights, intuitions, emotions, mental models, values, and actions are some of the tacit knowledge.

Explicit knowledge: can be easily expressed in words/numbers and shared in the form of data, scientific formulae, product specifications, manuals, and universal principles. It is more formal and systematic form of knowledge.

Heuristic knowledge: the Heuristic knowledge is the specific rule-of-thumb or argument derived from experience and learning.

Commonsense knowledge: is the general purpose knowledge expected to be presented in every normal human being. Common-sense ideas tend to relate to events within human experience.

Metaknowledge: can be defined as knowledge about knowledge. It is knowledge about the operation of the knowledge based system i.e. about their reasoning capabilities.

Domain knowledge: is valid knowledge to be applied to the specified domain. Specialists and experts develop their own domain knowledge and use it for problem-solving. The Knowledge acquisition is a bottleneck for the development of KBS because, it's trustworthiness and the performance mainly depends upon reliability, validity, and accuracy of the acquired knowledge [51]. On the other hand, acquiring the knowledge from the experts cannot be easy due to the following difficulties [18].

- The experts may not tell all what they know to the KBS builders or knowledge engineers and may not show consistency in behavior when they are observed or interviewed,
- Experts may not be willing to or have time scarcity for eliciting knowledge and may not know how to communicative their knowledge or even may not able to do so.
- System builders tend to gather knowledge from one source, but the relevant knowledge is distributed across several sources. And the knowledge collected may not be complete due to builder's attempt to collect from documented knowledge are some difficulties to elites knowledge from experts.

Therefore, the knowledge elicitation merely affects the construction of the KBS, if not used properly the combination of different techniques to overcome the difficulties raised above. Thus, to eliminate these difficulties [52] suggested the following methods that are commonly used knowledge acquisition techniques.

a. **Interview** is the process of direct interaction with the domain expert on how they perform their tasks based on their expertise. This is can be procedural knowledge acquired from direct elicitation methods. Depending on its structure, interview can be classified into the structured or goal-oriented questioning of domain expert by knowledge engineer directly,

semi-structured or flexible with both closed-ended and open-ended questions, and unstructured interview conducted informally and provide well organized description of the cognitive process [53].

b. **Observation**: is the process by which the knowledge engineers directly observed how the domain experts solve the problems by observing what the experts' physically do when solving the problems.

c. **Document analysis**: to collect relevant and detailed knowledge from existed documents in different format. These documents can be professional literature, brochures, manuals, guidelines, employee handbooks, reports, glossaries, course texts, and other relevant materials for the study.

d. **Data mining**; the Knowledge discovery by data mining is a process that extracts implicit, potentially useful and previously unknown information from the data. Data mining and machine learning are the main approaches for automatically knowledge acquisition. These are more advantageous than other methods of knowledge elicitation as it generates the knowledge from available data. Acquiring the knowledge from domain experts is difficult since the human knowledge is tacit, domain experts are uncooperative sometimes and usually busy and good knowledge engineers are also expensive. Thus, Abdulkarim [18] set the following as some objectives of knowledge acquisition automatically.

- To increase the productivity of knowledge engineering and reduces the cost of manual knowledge acquisition.
- To reduce drastically or even eliminate the need for an expert for acquiring the knowledge and To increase quality of the acquired knowledge

To accomplish the above mention and other advantages of acquiring the knowledge automatically the above discussed four knowledge acquisition methods i.e. interview, observation, analyzing documents, manuals and guidelines and data mining classification techniques have been applied for this research work to acquire better knowledge.

➢ Knowledge Representation

In artificial intelligence, once the knowledge has been acquired using different techniques it should be represented into the computer system in a way human can understand to solve problems. Knowledge representation in artificial intelligence can be expressed as a way of representing a validated knowledge collected from experts or induced from available set of

data in an understandable way for human experts and executable on computers. It is the systematic encoding and transfer of human expert's knowledge in an appropriate way [54]. The production rule, Frames, Decision trees, Objects, and Logic are the most popular knowledge representation methods.

Production rules: knowledge represented in the form of antecedent-consequent pairs. A Rule is a structure containing IF and THEN components. Rules can be viewed as the simulation of intellectual behavior of human experts. IF condition or premises occurs THEN action or consequence will occur. This can be represented as

IF <condition> THEN <action>

Production rule represents knowledge in notational suitability with easy expression and declarative form of knowledge representation. This allows the knowledge base to be expanded by adding more rules. These are easy to understand, criticize and improve.

Decision trees: simplified knowledge acquisition process and easily converted to rules. The conversion can be accomplished by a computer program.

Predicate calculus: predicate logic is to indicate relationships and their reasoning. Symbolic logic systems are used to represent rules and procedures for a computer to perform reasoning using logic to draw inference from premises. Predicate logic is a basis for Prolog (Programming in Logic) and provides the theoretical foundation for rule-based systems.

Knowledge Base: Knowledgebase is the vital component of a knowledge-based system and set of sentences expressed in knowledge representation language. Each sentence existed in the knowledge base are a declaration about the real world. Knowledgebase consists of a set of rules called the Rule Base and data or facts called the database. The rule base and the database of a KBS together are called the Knowledge Base and both are created by knowledge engineers and contain problem specific knowledge [50].

Inference Engine: Inference engine is the software program set to refer and manipulates the existing knowledge and makes decisions about actions to be taken and to mimic the reasoning ability of human expert arriving at the conclusion. Inference engine simulates the evaluation process relating to the information and rules in the knowledge base to the answers to a series of questions asked. Further, it deduces facts or draws conclusions from the knowledge base by using pattern matching and searching techniques. In addition to

examining existing rules and facts, it also adds new facts and infers new knowledge as needed by the users [50].

2.3.5. Knowledge Validation

In building KBS Knowledgebase testing structures can be adapted to validate the set of rules generated from data. The discovered knowledge needs to be validated for its accuracy, consistency, and completeness. The errors in rule-based knowledge bases classified into two; consistency errors (includes redundant, conflicting and subsumed rules) and completeness errors (which includes unreferenced attributes, illegal attributes, and unreachable rules) **[**21]. Further, the author set approaches to validate knowledge bases discovered from databases since the knowledge discovered from databases is different from that of expert system in the way it created, which in return affects the way it validated. However, because of availability and simplicity discovered knowledge based on domain knowledge is has been employed for validation of knowledge for this research.

Domain Expert: Domain expert is a person who is an expertise in his/her domain area. For example, a medical doctor who gives medical aid for patients and development agent who gives the advice for farmer and farm institutions are domain experts.

Knowledge Engineer: Knowledge engineer is someone who gathers knowledge from experts through interview or using automatic knowledge acquisition techniques. The knowledge engineer needed to have the knowledge of a Knowledge Base System development technology and know-how to develop it using a development environment. It is not necessary that the knowledge engineer be proficient in the domain area in which the expert system is being developed, but the general knowledge and familiarity with key terms is desirable [50]. Similarly, Knowledge engineers are the professionals who know how to capture knowledge from the domain experts and structure it in a form that can be processed by the computer and computer-based expert systems [49].

2.3.6. Knowledge Based Reasoning Techniques

Depending on the application and usage reasoning of the Knowledge Base System different reasoning techniques i.e. Rule-Based, Case-Based Reasoning, Fuzzy Logic, Semantic Network, and Neural Network can be applied.

Case-Based Reasoning (CBR) is defined as "the adapting old solutions to meet new demands, using old cases to account for new situations, using old cases to evaluate new solutions, or reasoning from precedents to interpret a new situation". It is more suitable to make better Decision in dynamical environment as it reuses the previously solved problems and learning from experiences for future decision. It is advantageous in its Modularity, Easiness of knowledge acquisition, and self-updatability. It also limited in its Inability to express general knowledge, Knowledge acquisition problem and Inference efficiency problems and Provision of explanations to the drawn solutions are some disadvantages of using case-based reasoning [54].

Rule Based Reasoning: the acquired knowledge is represented by set of rules and facts. It's most popular mainly due their naturalness, and simplicity which facilitates comprehension of the represented knowledge. The knowledge basically represented in the form of rules, or If<condition> then <conclusion> format. Logical connectives such as, AND, OR, and NOT are used to connect condition of rules. When necessary conditions of rules are satisfied, then the conclusion is derived and the rule is said to be executed or fired. Rule based expert systems are applied in different domain areas such as medical diagnosis, electronic troubleshooting, and data interpretations [55].

- **Rule-Based Reasoning Techniques**

Inference engine of Rule-based reasoning used forward chaining and backward chaining strategies to draw the conclusion from the knowledge base [55].

Forward Chaining is the data-driven technique in which the system starts with the initial set of elements or available information in the working memory and keeps on firing rules until there are no rules to be applied or the goal has been reached or conclusion drawn. Consequently, the system is moving forward from the current state to a goal state and the problem can be split into sub-problems and solved. This reasoning strategy is suitable for applications in which the number of goal states is large.

Backward Chaining: Backward chaining is a goal-driven approach/ method which starts from an expectation of what is to happen (hypothesis), then seeks evidence that supports (or contradicts) expectation. This entails formulating and testing intermediate hypothesis (or sub-hypothesis). Backward chaining is used when the number of possible initial states are large, compared to the number of goal states. The tasks of classification and diagnosis are best

suited for backward chaining [55]. Rule-based reasoning approaches are advantageous for development of Knowledge Base System. The Compact representation of knowledge, homogeneity, and independence, the naturalness of representation and modularity or discreteness of knowledge unit are some advantages of rule-based reasoning techniques [51]. In contrast, Rule-Based Reasoning is limited as it becomes Bottleneck to the Knowledge acquisition, Brittleness/fragility of rules, Inference inefficiency, Interpretation problems and Difficulty in maintaining large numbers rules are some limitation of rule based reasoning approach [49].

2.3.7. Application of Knowledge Based Systems in Agriculture

Knowledge Base System can be applied in identifying and solving problems in different domain areas where there is scarcity and readily unavailability of human experts. The KBS applied in agricultural domain area as the system for diagnosing and treating the animals diseases, plants diseases and various crop pests, chemical and pesticides managing, weather damage recovery, and mixed applications in helping agriculture researchers, farmers and development agents [56]. Many scholars recommended application of knowledge based systems to overcome the problems of agriculture. Example [52] specified the main areas of KBS application as the following.

Crop Management Advisors: this mainly helps the farmers by providing decision support for the process of growing crops and produces recommendation about usage of fertilizers.

Pest Management and Diagnostic System: helps the farmers to identify problems of pests and gives advice for managing the problems. This can be system for integrated pest management like apple orchards. The diagnosis concerned with any kind of diseases in plants and any crops. It is domain specific and generic tool for diagnosing plant diseases.

Livestock Management Advisors: this system helps to give integrated advice for animal breeders.eg. Application of conditional causality in an integrated KBS for daily farms is one module. The system also constitutes health, production, and financial modules. It gives decision support for daily farm management tasks.

Marketing Advisor Systems: this helps the farmers by providing advice to on marketing of different farm products. These include Cattle and Grain Marketing which helps farmers to select different marketing alternatives for their cattle, and grain. On the other hand, the

systems like process control, conservation or engineering systems and planning systems are also applied in agriculture.

2.4. Review of Related Works

Many studies have been conducted for exploring the application of data mining and Knowledge Base Systems in agriculture. The previous research attempts in the agriculture area have realized the applicability and significance of Data mining and Knowledge Based System. Most of the studies on the application of Data mining in agriculture used data mining technique to predict future crops production, weather forecasting, pesticides and fertilizers management, and revenue to be generated from agricultural yields [56]. Many authors applied different data mining techniques to improve its applicability in agriculture. For example, Yuan et.al [39] applied Bayesian Network to develop data mining model to analysis the agricultural land gradation data. They experimented on dataset containing 2000 cases and developed model with accuracy of 87.5% compared with Naive Bayes and decision tree and concluded it provides scientific evidence in decision making. Although, Wang et.al [57] Applied Support Vector Machine (SVM) to predict Wheat crop disease.

In a similar way, Yang et al [58] investigated to apply data mining predictive models to the agriculture data to predict and analysis occurrence of Wheat crop stripe rust by building Bayesian Network model. Rahul et.al [40] applied Data mining techniques to predict the annual yield of major crops and to recommend planting of different crops. The objective is to estimate better crop yield for Indian districts. In this work clustering and classification, tasks have been applied. Thus, Districts are clustered into distinct cluster containing the similar attributes. Accordingly, they categorized into four clusters based on environmental or climatic attributes, biotic factors, irrigated area and individual crop yields of rice, potato, and Wheat. K-means clustering algorithm from Rapid Miner studio was applied for clustering and linear regression, k-nearest neighbor, and neural net algorithms have been applied to classification models.

Therefore, the above discussed and other research conducted in agriculture by applying data mining their works limited only in analyzing the data and lacks comparison of results of different mining algorithms instead applied one technique per research only. It also lacks integration of Data mining results to build problem-solving systems and support decision making like expert systems. Similarly, most of the local researchers focus on application of data mining in health and left agriculture which is the backbone of the country's economy. On the other hand, the Knowledge Base Systems with capability to pinpoint agricultural

problems and figure out how to solve it has been implemented by different researchers. Ejegu [14] conducted research on cereal crop diseases identification and treatment by developing the knowledge based system prototype that assists agriculture experts to make timely decisions on cereal crop disease diagnosis and treatment. The knowledge was acquired through interview and document analysis. Mainly five domain experts were interviewed to elicit knowledge. Detailed knowledge acquired on symptoms, treatment, and identification of disease and modeled using decision tree as cause and effect relationship between symptoms and pathogens that cause diseases. The Swi-prolog programming and rule-based knowledge representations were used for implementation. This work lacks including data mining predictive models to predict determinants and patterns of disease to take proper protection and control of diseases. The research limited to analyze and represent the knowledge of five experts only which may not result in adequate knowledge for knowledge-based construction.

Similarly, Saad et.al [59] investigated to develop a web-based expert system to diagnose diseases and pests of Wheat crop in Pakistan. The e2gLite expert system shell with Java interface has been used as development tool. The detailed survey and interview of a farmer, experts, and review of related articles to applied to acquire the knowledge. The ruled based knowledge representations were used to represent knowledge.

Locally Feidu [60] investigated to develop web-based bilingual agronomy advisory system for Wheat and maize crop production. The developed system prototype intends to support extension workers and researchers. Interviewing experts and document analysis used to acquire the knowledge. Their work lacks work also lacks automated knowledge acquisition.

Generally, all the above discussed and other local researches conducted in the agricultural area intended to develop the Knowledge Base System. But, the research attempt lacks mining and analyzing the agriculture data. It also lacks integration of Data mining results with the knowledge base systems and limited to acquire knowledge manually by interviewing experts; which results in insufficient knowledge for building Knowledge Base System.

Abdi [61] explored the problems of absence in a comprehensive and integrated knowledge of fraud risk management in insurance companies. The research applied and recommended automated inducing of hidden knowledge from the data. Furthermore, stated the necessity of Knowledge based system in solving complex problems across with a variety of approaches based on the knowledge representation method. The research applied automatic knowledge assimilation method to build Knowledge Base System. Rule-based reasoning approach for knowledge representation has been employed, because of it's the simplest, popular,

expressive, and easy to interpret, generate and to use. The SWI-Prolog programming tool was applied to build a knowledge base system. The researcher attempted to integrate data mining obtained results with knowledge based system for fraud detection. Clustering algorithm applied for grouping insurance claims and followed by classification for developing the predictive mining models. The PART classifier algorithm was applied to the dataset for generating rules. Integrator application created with Java and linked to PART and prolog.

The aim is to design the Knowledge Base System that identifies fraud claims by using an automatically constructed knowledge base. It performs identification based on the knowledge acquired from Data mining predictive models and provides advice for claim analyst. This research work mainly limited to Use only the automated knowledge acquisition method from predictive model and their prototype also lack attractive user interface. It also limited to use prolog interface since accessed by command line interface. In a similar way, Abdulkarim [18] has been tried to integrate mining predictive model induced knowledge with the knowledge based system. The study selected using the KDDcup‘99 dataset from ACM Knowledge discovery site. The research tried to construct rule-based intrusion detection and advising knowledge based system by applying hidden knowledge extracted from the dataset. It used rule-based induction algorithms for data mining. The JRip classifier algorithm selected and applied to the dataset. The researcher employed the classification technique and decision tree algorithms to detect network intrusion. This work also lacks addition of knowledge acquired from domain experts and the prototype lacks graphical user interface. Tesfaye [62] also conducted research in human health care with the objective to explore the opportunity of integrating knowledge bases system with data mining techniques for diagnosis and treatment of diabetes patients. The Prolog programming language for knowledge representation and Java used for constructing integrator application. This prototype also lacks graphical user interface that makes simple usability.

Therefore as to proposed research, no research has been conducted to build an integrated predictive model and the knowledge based system for Wheat disease detection and treatment at Arsi Wheat production area. Having this the proposed research will fill the gaps by building integrated predictive mining model and knowledge based system for supporting detection and treatment of Wheat Diseases. Finally, the main objective of this research is to build the predictive model to extract hidden knowledge and integrate with Knowledge Base System for detecting and treating Wheat diseases that could help in the effort on enhancing decision.

CHAPTER THREE

METHODOLOGY

3.1. Research Area/Setting and Period

This study was conducted at Arsi Zone Wheat production areas. It focused mainly Kulumsa Agriculture Research Center specifically on Wheat pathogen division for data and expertise support. The study was started understanding and analyzing domain area to collect data and necessary information needed to conduct it; from February 2017 to January 2018 for implementation and evaluation of the results.

3.2. Method of Study

To achieve the objectives of this study the following research methods and techniques have been employed. This study intended to use sampling technique and knowledge discovery process models as method of study. Thus, Purposive sampling technique has been applied to acquire the knowledge from the domain experts; since it is common technique in qualitative research where participants decided to preselected standards appropriate to specified question. For this purpose domain experts have been selected for interview based on profession specialization, educational qualification level, year of experience and immediate position in wheat pathogen department.

The study design set to use mixed research approach i.e. the knowledge discovery process models for accomplishing data mining objectives, sampling techniques for knowledge acquisition from domain experts and prototyping system development approach. The knowledge discovery process model has been used to design the research since it is a nontrivial process to identify valid, novel, potentially useful and ultimately understandable patterns from large collections of data. There are many knowledge discovery processes models designed and developed by different scholars. [74] Made Comparison among models developed by Fayyad [28], Cios [29], Anand et.al [63] and Cabena et.al [64] and concluded that the hybrid knowledge discovery process model with six step of KDP was successfully implemented saving the project cost and time. Considering this conclusion and appropriateness of the model this study, we have implemented hybrid knowledge discovery process model. Correspondingly, the model includes one step as an initiator for other step and is the iterative and interactive model. Therefore, from experience of successfulness of its

implementation, in academic projects and this study also conducted for the academic purpose, CIOs et.al [29] six step hybrid KDP model has been implemented for this study. Therefore the main justification for this study to use this model is the research conducted for academic purpose and the model contains simple and short listings of specific steps that lead to knowledge discovery using Data Mining techniques. It also initiated the research starting from understanding the business area to discover and use the knowledge; of which the Data mining is the core to be performed following other successive steps. Therefore the study implemented the hybrid KDP.

3.3. Data Collection and Knowledge Discovery

The appropriate source data for conducting this study has been determined. Thus primary and secondary sources of data and knowledge identified. The primary source of data for this study is domain experts. To elicit the required knowledge from the domain experts' interview data collection method has been applied. For this purpose purposive sampling has been employed since it is the main way to elicit view of experts who have knowledge about specified domain area. Therefore, it is appropriate way to capture demonstrable experience and knowledge of experts. The data and knowledge also acquired from secondary source by review and data mining techniques. Data mining techniques has been undertaken to acquire the knowledge from selected dataset. Therefore, the process of knowledge acquisition for this research work contains all the basic activities of collecting the required data and analyzing, identifying important concepts focusing on causes and symptoms of Wheat Diseases. The acquired knowledge contains both primary and secondary sources of knowledge. The basic techniques and activities involved in this study to collect data and extract relevant knowledge from data and domain experts are the following.

3.3.1. Reviewing Relevant Documents And Manuals

This technique helps to acquire the explicit knowledge from the existing documents, manuals and guidelines, books, and internet as secondary sources of knowledge. While conducting this study the related documents, manuals and guidelines were reviewed and analyzed. As a result of document analysis relevant and technical knowledge were extracted and added to knowledge acquired from predictive models.

3.3.2. Interviewing Domain Experts

In this study Interview has been undertaken to elicit tacit knowledge from domain experts. For this purpose, both structured and unstructured interviews were used to elicit the

knowledge. Key informant from the research area was sampled and interviewed. These are agriculture researchers and experts working with different educational background and experiences. Thus, two researchers working at Wheat pathology department in kulumsa Agriculture research center, two lecturers and researchers selected from Madda Walabu University School of agriculture and two agriculture experts or development agents were interviewed for this study. In general total of six plant science experts, Wheat specialist and pathologists have been participated in the interview to understand domain area, acquire additional knowledge and to identify the problems. These experts were interviewed about major diseases that affect Wheat and their symptoms as well as treatments undertaken to control. The obtained information recorded using all mobile phone record, paper sheet, and pen. Finally, the obtained data, information and knowledge analyzed and integrated to the results obtained from Data mining predictive models. The interview questions used for this purpose have been attached at Appendix III.

3.3.3. Data Mining Techniques

Data mining is the primary knowledge acquisition technique applied for acquiring the knowledge for this research work. It mainly used to build predictive model from selected dataset. For this purpose different rule induction algorithms and knowledge discovery tools can be used to analyze data and generate knowledge. Thus different rule induction classifiers algorithms have been applied for this research; since data mining algorithms were efficient techniques to extract the hidden knowledge from the huge datasets for decision-making activities [21]. Thus classification data mining techniques has been implemented for this study. Since the classical knowledge acquisition have own limitations, the integrated knowledge acquisition hybrid knowledge acquisition techniques were implemented for this study.

3.4. Knowledge Representation

In this study the knowledge acquired from different sources represented in a computer traceable form. For this purpose rule based knowledge representation has been implemented. Rule-Based knowledge representation method takes the form of or 'situation-action'. 'IF certain situation holds THEN a particular action should be taken. The production rule works in the form of “if the condition holds then the action is appropriate”. Therefore the rule-based knowledge representations are capable of modeling any computable procedure [65]. The Rule-based knowledge representation approach was applied to represent knowledge for this

study. The Rule-based knowledge representation is highly expressive, is easy to interpret and easy to generate. In addition, it also classifies new instances rapidly [18]. “It is prominent means to represent the massive amount of domain-specific knowledge in Knowledge based system”. To attain the advantage of rule-based knowledge representation; the rule-based knowledge representation has been implemented for this study.

3.5. System Development and Implementation Tools

For implementing the results the prototyping approach has been employed to develop knowledge based system since it allows involvement of domain experts and users for commenting on development and evaluating system performance and efficiency.

3.5.1. Implementation Tools

The essential step of the research design to achieve the objectives of study is selecting the appropriate research tools. Accordingly, the tools that help achievement of this research were selected based appropriateness, accessibility, and familiarity with the tools. Similarly, the nature of data to be mined considered containing labeled classes that are easier for applying classification data mining task. The nature of diseases identification and treatment by human expert also takes the form of rules, as well as data mining results also takes the form of rules. Thus these rules are also easily understandable by human experts. These all lead toward the implementation of rule-based or rule generating classifiers algorithms. Therefore, rule inducting classification algorithms are needed to be implemented. Since there is no specified standard that can be used for selecting classifiers more than the nature of research and the researchers’ consideration, this research selected and conduct the experiment with JRip, J48, PART, and REPTree. These are implemented through WEKA (Waikato Environment for Knowledge Analysis) since these algorithms are capable of generating rules from data. WEKA is powerful data mining tool developed at the University of New Zealand. The tool is written java, containing GUI and API or Application page interface for interacting with data directly. It also included many embedded java libraries that can be extracted and used in any Java applications [66].

Similarly, the software is platform independent, non-commercial, open source and includes powerful machine learning algorithms for data preprocessing and analysis. It also provides extensive support for whole process of data mining and accessed through a common interface for comparing input and output data.

Additional Microsoft- Excel, Java libraries like JavaFX, JESS and Eclipse have been selected and implemented. Thus Microsoft-excel used for storing and processing data set, since the initially collected data also store in excel format. Java libraries and Eclipse used for accessing and integrating rules generated in Data mining tools and designing the Knowledge Base System prototype. Further the results of each classifiers algorithm were compared and the best performer classifiers were selected in terms of accuracy and rule generating capability. These resulted in rules used for the Knowledge Base System construction.

3.6. Testing and Evaluation

Testing and evaluations of developed prototype has been undertaken extensively to measure its performance and issues of user acceptance. For this purpose test cases were prepared and evaluated both by system and domain experts to evaluate performance of prototype. User acceptance testing was also undertaken concerning issues of how well the system addresses the need of the domain experts in detecting diseases. To assess evaluators feedback questionnaires prepared from system evaluation standard and used. Evaluators interact with the system and evaluated the system by using closed ended questionnaires.

CHAPTER FOUR

EXPERIMENTATION AND ANALYSIS

The necessity for data mining technologies has been realized to analysis data and acquires knowledge from the large data since it is the solution for the problem of tacit-knowledge acquisition from domain experts by which the knowledge engineers are using traditional knowledge acquisition like interview and document analysis. However, the data mining is the knowledge discovery procedure to extract unknown and unseen knowledge and rules from plentiful data. The results of data mining predictive models can be made easily understandable and accessible by experts and non-experts. To make the process easier the Knowledge Base System has been applied.

4.1. Wheat Disease Diagnosis and Treatment Process

Before building data mining predictive models it important to understand how the domain experts detect and recognize the diseases. For this purpose understanding the business area and problems were undertaken at Kulumsa Agriculture Research Center which was established in 1966 by government of Ethiopia and the Swedish International Development Agency (SIDA). The Research Center is operated by Ethiopia Agriculture Research Institute and mandated to research on Wheat, malt barley, and highland pulse crops nationally. Also, it serves as Wheat Center of Excellence for East Africa (Ethiopia, Kenya, Uganda, and Tanzania) regionally. The analysis of documents and interview of experts resulted in the identification of common Wheat diseases that mainly affects Wheat in Ethiopia. These are leaf rust, stem rust, barley yellow dwarf, Fusarium, scald, take-all, Powdery mildew, smut, common bunt and net blotch. These interviewed experts responded that Wheat diseases diagnoses are mostly carried out using visual identification through observation crop parts. The Laboratory examination can also be undertaken in rare cases to identify the diseases.

To identify visually through observation, the symptoms of Wheat diseases are identified by experienced experts by applying both tacit knowledges accumulated through experience and explicit knowledge such as using keys/ colored photos of crop history and guidelines or principles of crop protection. The experts also stated that visual identification method is more advantages regarding time and cost saving than laboratory examination method. They also assured that since laboratory examination takes time, it cannot be used for providing an immediate solution for the outbreak of crops diseases. The experienced and knowledgeable

experts can easily identify the problems of crops through observation of symptoms and a sign of diseases that attack crop to make immediate decision, after particular pathogen that causes a particular disease is identified, the appropriate measure should be taken to control it. The experts used visual inspections to identify the diseases by seeing parts of crops as head/grain; stem, root, and leaf are merely affected by the diseases [14]. Additionally, some of the treatments to control Wheat diseases are to use resistant varieties, spray fungicides and cultural diseases control like crop rotations.

Further, the knowledge of disease epidemiology is crucial to design effective disease control measures, but few efforts have been made in the area of cereal crops protection mainly on Wheat and barley in Ethiopia. In addition, pathogenic fungi are by far the most important and yield limiting causes of diseases which attack these cereal crops. Of these diseases rusts (genera Puccinia), ago/smuts (Usti/ago), bunts (Tilletia), powdery mildews (Erysiphe), Septoria, Alternaria, Helminthosporium, and Fusarium Pythium are the most widespread, regularly occurring and potentially dangerous diseases throughout the world and have a devastating impact [67]. The Wheat production in Ethiopia was mainly constrained by both technical and socioeconomic factors. Further, the incidence of the different Wheat diseases varies with altitude and other climatic factors i.e**.** Stripe rust disease is highly prevalent to high altitude and cooler areas. But now it has extended its adaptation to lower and warmer altitude areas [68].

Wheat Rusts: The rusts are common Wheat crop diseases suppressing productivity of Wheat throughout the world. These are categorized as leaf rust (caused by P. triticina), Stem rust (caused by Puccinia graminis f. sp. tritici), and stripe rust (caused by P. striiformis) are some of the most widely recognized diseases of Wheat with their respective pathogens. Leaf rust and Stem rust disease occur almost in all area wherever Wheat is grown whereas stripe rust is generally restricted to areas of cool temperature in the growing season [68].

Leaf Spotting: can be the Septoria, leaf spot and Tan spot are the two most common Wheat crop leaf diseases caused by the fungus. The Symptoms are both form very small, yellow spots on the leaves when the fungus first infects the leaf and expand into distinctive patterns of the plant. The knowledge elicited from different sources mainly focused on Wheat diseases, its symptoms and types of treatment measures undertaken. The analysis of domain experts' interview and evidence of dataset indicated that the diseases like Leaf rust, Stem rust, Stripe rust/yellow rust and Septoria tritici are the mostly invading diseases in the

selected research area. Therefore this study mainly focused and analyzed in detail since these are major crop diseases that affect the productivity of Wheat production throughout the country. Additionally, the details of types of Wheat disease and their symptoms have been attached to a table in Appendix I.

4.2. Data Understanding and Selection

Working early with domain experts, the problems, objectives of study and requirements from the business perspectives have been identified. After identifying and understanding the objective of the study, the knowledge of problem definition should be converted into Data mining objectives. This is to understand the business and its problems in setting the data mining objectives to solve these problems. Gathering and understanding data is the initial steps to discover knowledge from data. Therefore, the activities of understanding and selecting data were undertaken to start from identifying the data source.

➢ Data Source

The dataset for conducting this study was collected from Kulumsa Agriculture research Center. The center collected the data every production year for Wheat and other cereal crops as disease surveillance at Arsi agriculture area. The data was collected from farmers' field, experimental field, and state farm fields.

➢ Description Of The Initial Data

The data set collected for this study contains data for production years from 2014 to 2016. These were collected while surveying for outbreaks of Wheat disease throughout the selected research area. These data collected by different researchers. By discussing and working with the domain experts we have collected the survey data taken by the researchers and experts during their fieldwork study. The survey data of the disease surveillance was taken every production "time of Autumn, Summary, and Winter" or the Main season. This survey was taken early as crops sawed, in the mid time and early before the harvest time. The production times of main season were similar performed by all farmers and state farms throughout the country and selected for this study as suggested by domain experts. This is because the production time of main season performed by all farmers and state farms in a similar way. The dataset contains 4709 records/instances with 20 attributes and 4 classes of diseases as Leaf rust, Stem rust, Yellow rust and Septoria. The center collected data for the research purpose as diseases' surveillance for each production years, from the farmers' field, state

farm and from the experiment area. The data filed and visualized using standard Microsoft Excel sheet. But the initial data set is not set in a format that is ready for data mining purpose. Table 4.1 generalizes the description of the initial dataset.

Division	Total
Number of instances	4709
Number of Attributes	20
Number of classes	4

Table 4-1 Descriptions of Initial Data Set

Since the selected data set is not in the format suitable for mining, preprocessing should be performed on it. This is to verify the data quality for the missing values, error values and to obtain high-level information regarding the data mining questions to be answered. To make data suitable for data mining the [69] clearly stated the five points to be considered to make data in a particular format.

- ✓ First, all the dataset to be mined should be in a single table;
- ✓ Second, each row of data should correspond to an entity that is relevant to the business activities.
- ✓ Thirdly, the columns of data with a single value should be ignored from the dataset.
- ✓ Fourth, the columns of data with a different value for every column should be ignored from the dataset.
- ✓ Finally, for building predictive models, the target column should be identified and all synonymous columns should be removed.

On the other hand, attribute selection should be performed due to some mining algorithms slow with too many attributes and some attributes may not be relevant for analysis. Therefore, taking the above points into account and working with domain experts' relevant attributes for predicting the Wheat diseases has been selected. For these purpose WEKA standard attribute, filtering methods were applied. These resulted in removed irrelevant and redundant attributes for faster model building. Here table 4.2 shows the description of selected attributes.

No.	Name of Attributes	Data Types	Description
1.	Year	Numeric	Describes the year in which the data has been collected. Contains 3 distinct numeric values i.e. 2014,2015,and 2016
2.	Date	Nominal	Describes the Specific date at which the disease survey was taken from fields. Contains five distinct nominal values in months: i.e. August, September, October, November, and December,
3.	District	Nominal	Describes the specific research area from where the data were collected. Contains 22 nominal distinct values i.e. G/Asasa, Nagele, shaashemene, Kofale, Dodola, Adaba, Tiyo, Hetosa, Lodehetosa, Diksis, Arsi robe, Ticho, Kula/Sude, Sire, Chale, Guna, Marti, Gonde, Itaya, hunkolowabe, Lemu-bilbilo, Digalu,
4.	Altitude (m.a.s.l)	Numeric	Describes the measured height of a place above the sea level. Contains 115 numeric distinct values.
5.	Previous Crop	Nominal	Describes the crops which have been sawed on the fields before the season the survey was taken. Contains 7 nominal distinct values i.e. Wheat, maize, barley, potato, fallow, fababean, and maize and teff…
6.	Variety	Nominal	Describes the types Wheat crops adapted in the area. Contain 18 nominal distinct values i.e. Kakaba, Kubsa, Danda, Digalu, kubsa+mixture, Paveen, digalu+mixture, K62954A, Triticale, Israel, unknown, ET13A2, maddawolabu, Hulluka, Hidase, Ogolcho, enkoy, and shorima …
7.	Growth stage	Nominal	Describes the series of steps passed by the crop beginning from saw to harvest. Contains 14 nominal distinct values i.e. Dough, Early-dough, Soft-dough, Hard-dough, heading, Early-flowering, Flowering, flowering-complete, booting, Milky, Milky–soft-dough, milky-flowering, Tillering complete and maturity.

8.	**Field condition**	**Nominal**	**Describes the general condition of the field from which the survey was taken. Contains 3 nominal distinct values i.e. good, moderate and bad.**
9.	Soil Type	Nominal	Describes the Types of soil from the fields of the surveyed area. Contains 5 nominal distinct values i.e. light(sand), Black(heavy), Red(clay), Red and black.
10.	Drainage Condition	Nominal	Describes the condition of the water is able to flow in soil. Contains 3 nominal distinct values i.e. good, bad and moderate.
11.	Favorable conditions	Nominal	Describes the condition of temperature which is favorable for the development of Wheat rust. Contains 3 nominal distinct values i.e. cool/wet, low/wet, moderate/wet
12.	Parts of crop affected	Nominal	Describes the parts of crop affected by the diseases. Contains 3 nominal distinct values i.e. stem/leaf-sheath, leaf-blade, leaf-sheath/spike.
13.	Degree of damage	Nominal	Describes the degree of damage seen on parts affected by the crop. Contains 3 nominal distinct values i.e. tearing outer parts, seed-shriveling, non-tearing.
14.	Lesion color	Nominal	Describes the colors of lesions observed in the affected parts of crops. Contains 4 nominal distinct values i.e. reddish-brown, orange-brown, tan and orange.
15.	Lesion shape	Nominal	Describes the shapes of lesions observed in the affected parts of crops. Contains 3 nominal distinct values i.e. oval-elongated, elongated-blistered and round-narrow-stripe.
16.	Disease type	Nominal	Describes the Disease types are the divisions of the Wheat crop rusts. Contains 4 nominal distinct values i.e. yellow rust, leaf rust, Stem rust and Septoria.

Table 4-2 Descriptions of Selected Attributes

Explanatory Data Description

The main objective of exploratory data analysis is to investigate the variables, examine its distributions in data and the relationships among them. These explanatory data description can be applied to examine the distribution of the response variables and its relationships with predictor variables and to explain their relationships. These relationships measure the strength of dependency among attributes. Therefore while conducting this study, being with domain experts the relevant attributes from the dataset were selected containing 16 attributes of which 15 were predictors or independent variables and one dependent variable.

Here table 4.3 shows a summary of predictor variables with its distinctive values.

Categorical attributes	Number of unique values			
Date	6			
District	22			
Previous Crop	7			
Variety	18			
Growth stage	14			
Field condition	3			
Soil type	5			
Drainage condition	3			
Favorable condition	3			
Parts affected	3			
Degree of damage	3			
Lesion color	4			
Lesion shape	3			
For continuous numeric values		Mean	S. deviation	Range
Year	3	2015.279	0.842	2
Altitude	115	2413.1	235.519	1113

Table 4-3 Predictor Variables with Number of Distinct Values

The dependent variable is a categorical variable called Disease types containing four classes as Yellow rust, Stem rust, Leaf rust, and Septoria. From the table above the proportion for each types of Wheat crop diseases shows that for the selected data sets the class of dependent

variable Septoria dominates the area covering 35.68% of the disease occurrence at the area whereas the Stem rust, Leaf rust, and Yellow rust covers 28.75%, 18.33%, and 17.24% respectively. The distributions of the dependent variables throughout dataset are shown in Table 4.4.

Category	Counts	Percentage
1. Stem rust	1354	28.753%
2. Septoria	1680	35.676%
3. Leaf rust	863	18.326%
4. Yellow rust	812	17.243%
Total	4709	100%

Table 4-4 Distribution of Dependent Variable

4.3. Data Preprocessing

Data preprocessing is the techniques of preparing the data for mining tools by removing inconsistent, redundant and noise data. These can be filling missing values and combining attributes and integrating multiple sources of data. Handling Missing Values are the most important step in data mining because the presence of missing values in a dataset can affect the performance of a classifier constructed using that dataset. Therefore the literatures stated that the Availability of less than 1% missing data is generally considered as trivial and 1-5% can be manageable. However, 5-15% of missing data values requires some sophisticated methods to be handled. Consequently, more than 15% of missing data may severely impact any kind of data analysis and interpretation made with it [70]. Several methods have been suggested in data mining and related literatures for treating these missing data values. These include filling the missing value manually, replacing all missing attribute values by the same constant, using the most probable value to fill the missing values which may not be visible and time consuming approach and also, replacing a missing value with the attribute's mean or mode value (for numeric or nominal attributes, respectively).

This last approach is the most commonly used method of handling missing values in a dataset [70]. The Wheat data used for this study contains the missing values that are due to an incomplete report of survey data. The missing values in data set are mostly left blank. Since the missing values of the data are not more than 5-10% which needs sophisticated methods for handling missing values, WEKA is tolerant to missing values, and it can be manageable. All attributes with missing values among the selected attributes are nominal attributes.

Therefore WEKA used to replace missing values with attribute's modal values for nominal attributes and attributes mean values are used for the numeric attributes. Here is the percentage of attributes with missing values shown in Table 4.5

No.	Attributes	Count of missing values	Percentage of missing values
1.	Date	10	0.212%
2.	Previous crop	11	0.233%
3.	Growth stage	29	1.0%
4.	Soil type	47	1.02%
5.	Drainage condition	33	1.0%
6.	Favorable condition	247	5.0%
7.	Parts affected	170	4.0%
8.	Degree of damage	245	5.0%
9.	Lesion color	284	6.0%
10.	Altitude	19	0.403%

Table 4-5 Percentages of Attributes with Missing Value

For this study the attribute evaluator GainRatioAttributeEval which evaluates the worth of an attribute by measuring the gain ratio with respect to the class has been applied; since it is the fastest and simplest ranking method. The Ranker search method was selected and used to rank attributes in order. The features or attributes with adequate scores were selected and ranked as important as well as the attributes with inadequate scores eliminated from ranking. Therefore the attributes with the higher gain ratios were selected and used for this study. According to [69] the columns with the similar attribute values or single-valued should be ignored. Accordingly, the columns of the selected data set like region, zones, Wheat type and season type; contains respective values Oromiya, Arsi, bread Wheat and the main season has been ignored from the data and does not use for building the predictive model hence they are relevant. Figure 4.1 illustrates the attribute ranking in order of gain ratio.

```
Search Method:
        Attribute ranking.

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

Ranked attributes:
 0.889   15 LesionShape
 0.866   12 PartsAffected
 0.842   14 LesionColor
 0.842   13 DegreeofDamage
 0.787   11 FavorableConditions
 0.373    4 Altitude
 0.353    3 District
 0.336    6 Variety
 0.324    9 SoilType
 0.261    7 GrowthStage
 0.244   10 DrainageCondition
 0.175    2 Date
 0.116    5 PreviousCrop
 0.105    8 FieldCondition
 0        1 Year

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

Figure 4-1 Attribute Ranks

Data preprocessing ends cleaned data by replacing missing values and removing redundant instances, which resulted in reduced data size of 16 attributes, 4709 instances, and 4 classes. Further dealing with domain experts' columns with most missing values also ignored. The columns with mostly missing attribute values were totally ignored from the table. This resulted in four classes of disease types. Since data were at a different table by integrating the tables most of the classes with almost missing values also ignored. To make dataset suit for data mining it should be converted into a comma delimited excel or comma separated (CSV) format, then to attribute relation file format (ARFF). For this study WEKA version, 3.8.1 applied. Since WEKA can load both CSV and ARFF file formats first file in spreadsheet saved as CSV and opened in WEKA explorer then saved as ARFF. The ARFF file contains three parts top header contains the name of relation to @RELATION, list of attributes or the column in data and their type as @ATTRIBUTE and data as @DATA.

This ARFF format can give data and does not show specifically which attributes is supposed to be predicted from the data as well it can also be used find association rules or clustering and any. Figure 4.2 shows the ARFF sample of the data set.

```
@relation 'Wheat Data'

@attribute 'Year ' numeric
@attribute Date {October,November,'September ',December,August,September}
@attribute District {G/Asasa,Nagele,Shashemene,Kofale,Dodola,Adaba,Tiyo,Hetosa,'Lode
@attribute Altitude numeric
@attribute PreviousCrop {Wheat,Maize,Barly,Potato,Fallow,'Faba bean','Maize and Teff
@attribute Variety {Kakaba,Kubsa,Danda,Digalu,Kubsa+Mixture,Paven,Digalu+Mixture,K62
@attribute GrowthStage {Dough,EarlyDough,Milky,Floweringcomplet,Flowering,EarlyFlowe
@attribute FieldCondition {Good,Moderate,Bad,'Good '}
@attribute SoilType {Light(sandy),Black(heavy),Red(clay),Black,Red}
@attribute DrainageCondition {Good,Moderate,Bad}
@attribute FavorableConditions {Moderate/Wet,Cool/wet,'Low,wet'}
@attribute PartsAffected {Stem/L-sheat/blader,L-sheat/spike,'L-blade,sheath'}
@attribute DegreeofDamage {Tering-outer-parts,Seed-shrivelling,Non-tearing}
@attribute LesionColor {Red-brown,'Tan ',orange-brown,Orange}
@attribute LesionShape {Oval-elongated,Elongated-blister,Round-narrow-stripe}
@attribute DiseaseType {StemRust,Septoria,LeafRust,'StemRust ',YellowRust}

@data
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
2014,October,G/Asasa,2372,Wheat,Kakaba,Dough,Good,Light(sandy),Good,Moderate/Wet,Ste
```

Figure 4-2 Sample ARFF Files

4.4. Experimentation

A total of the four experiments aimed at building predictive models were undertaken. The data collected for experiment contains all attributes and features, as well as all selected attributes involved in all experiments. After preprocessing the data, transforming and preparing in a suitable format for mining next activity is to undertake an experiment. Then selected classifier algorithms applied to build predictive models. For the experimentations; PART and JRip selected from rule induction classifiers as well as J48, and REPTree selected from tree-based classifiers and applied for undertaking the experiments. For each experiment, the default values of parameters in WEKA classifiers algorithms were considered.

4.4.1. Building Predictive Models

Four experiments with respective classification algorithms have been undertaken to build predictive models. The predictive models were built with commonly used supervised machine learning classification and prediction techniques that are decision tree inductions and rule induction algorithms. In general, the selected classifiers' algorithms are capable of

generating rules from a dataset which helps to understand hidden knowledge in data and later used for supporting decision making.

Experiment 1 – PART Classifier

The PART classifier uses the separate-and-conquer method to generate its' decision list from the classes. The classifier builds partial C4.5 decision trees for each iteration to make the leaf into rules. For this experiment the classifier generated 17 rules by involving all selected attributes of the dataset and 10- fold cross-validation test option with default parameters values were used. In this experiment from 4709 instances, 4684 are correctly classified by achieving the accuracy of 99.41% and Incorrectly Classified 24 Instances which are 0.59% of data and takes 1.7 seconds to build the model.

Experiment 2 – JRip Classifier

In JRip classifier the Classes are examined in increasing size and an initial set of rules for the class are generated using incrementally reduced error, JRip (RIPPER) proceeds by Advisement all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of the class. For this experiment the classifier generated 22 rules by involving all the attributes of the dataset and 10- fold cross-validation test option with default parameters values were used. Consequently, the experiment classified 4690 instances correctly by achieving the accuracy of 99.61% and Incorrectly Classified 19 Instances which are 0.39% of data. It also takes 0.42 seconds to build the model.

Experiment 3 – J48 Tree Classifier

J48 is popular decision tree generating and is an extension of Quinlan's earlier ID3 Algorithm. The algorithm can generate decision tree which can be used for classification and also referred as statistical classifier algorithm. 10–Fold cross-validation has been selected as test option with default parameters values. The tree with the size of 55 trees and 71 numbers of leaves are generated. The algorithm correctly classified 4680 which are 99.16 % and only 29 which are 0.84 % instances were classified incorrectly by taking 0.05 seconds to build the model.

Experiment 4 – REPtree Classifier

REPTree classifier is fast decision tree learner. It Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting) and only sort values for numeric attributes once. For this experiment 10–Fold cross-validation is selected as test option with default parameters values. The trees with the size of 283 were

generated. The algorithm correctly classified 4630 which are 98.80 % data and 79 which are 1.20 % of instances are classified incorrectly taking 0.17 seconds to build the model.

4.4.2. Models Evaluation and Selection Methods

As clearly mentioned in the above section two rule-based and tree-based induction algorithms have been used for the experiment. These selected algorithms allow generating rules from the dataset. The Results are evaluated based on the prediction accuracy of classification of instances into Yellow rust, Stem rust, Leaf rust, and Septoria. The prediction accuracy indicates the general classification accuracy of the classifiers algorithms. Therefore classifiers are evaluated to measure how they are correctly classified each class into their correct or incorrect class. Apart from the accuracy classifiers, the classification evaluation matrixes like True Positive rate, False Positive rate, Precision, Recall and F-measure are also used to evaluate and select the best model with the best performance. Table 4.6 illustrates the overall performance of the four classifiers based classification accuracy. Accordingly, the JRip has registered the best accuracy score with 99.61% of accuracy in terms of correctly classifying instances into correct classes and incorrectly classified 0.39% instances incorrectly. Further, the precision, recall and F-measure values of the classifier also evaluated and compared to other classifiers over all classes, which resulted in the best score for JRip classifier.

Classifier algorithms	**Correctly classified instances**		**Incorrectly classified instances**		**Time take to build the model(in second)**	**No. of generated rules**
	No.	Percentage	No.	Percentage		
PART	4684	99.41 %	25	0.59 %	1.70	17
JRip	**4690**	**99.61 %**	**19**	**0.39 %**	**0.42**	**22**
J48	4680	99.16 %	29	0.84%	0.08	71L/55t
REPTree	4630	98.80 %	79	1.20 %	0.27	283

Table 4-6 Performance of Classifiers

From the accuracy of classifiers stated above the JRip classifier takes more time to build the model when compared to J48 and REPTree with 0.08 and 0.27 seconds respectively. But the time taken to build the model cannot be taken as criteria to select the classifier alone. The

JRip model also generated 22 numbers of rules containing more features than another model in classifying instances. Therefore, the JRip classifier was selected in these objective evaluation criteria. The performance of classifiers also evaluated in terms of the error rate. This contains Kappa statistic which measures the accuracy of any cases to distinguish between reliability and validity of data. In comparison to other classifiers, the JRip classifier algorithm scored result of 99.3% in Kappa static. But in terms of another performance matric including Mean absolute error, Root mean squared error, Relative absolute error, and Root relative squared error its registered better error rates than others which means less in percentage. Table 4.7 illustrates the stratified cross-validation summary for each classifier.

	Classifiers	Correctly classified	Incorrectly classified	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Time taken to build model
1	PART	99.41 %	0.59 %	0.992	0.0055	0.0440	1.4179 %	13.383 %	1.70
2	**JRip**	99.61 %	0.39 %	0.993	0.0029	0.0410	0.5434 %	8.5958%	0.42
3	J48	98.16%	0.84 %	0.988	0.0083	0.0558	4.0304 %	16.9815 %	0.08
4	REPTree	98.80 %	1.20 %	0.984	0.0098	0.1024	10.986 %	31.1424 %	0.27

Table 4-7 Stratified Cross-Validations of Classifiers

On the other hand, the accuracy of classifiers was evaluated on each class by True positive, Precision, Recall and F-measure. Four experimented model registered scores with relatively little difference on all classes. Thus the JRip classifier scored True Positive rate of 99.9% on Stem rust making False Positive 0.001% which means near to 0%. The model also registered better True Positive rate for other three classes 0.997, 0.987, 0.996 for Septoria, Leaf rust and Yellow rust making 0.003, 0.013, and 0.004 False Positive rates respectively. The JRip classifier also registered 99.1%, for Stem rust, which means from the total Stem rust labeled 99.1% of the class actually found to be classified correctly. Therefore, the classifier scored

99.8%, 99.6%, 98.8% of precision for Septoria, Leaf rust and Yellow rust respectively. Similarly, it also actualized of total classes 100%, 99.5%, 98.7%, and 99.6% for Stem rust, Septoria, Leaf rust and Yellow rust respectively for each class. Hence the JRip classifier model registered better scores in terms of all Precision, Recall and F-measure and selected to be implemented for this study. Table 4.8 illustrates summarized accuracy by Class in terms of True positive, Precision, Recall and F-measure of classifiers with respective classes.

Classifiers		**Class**			
JRip		Stem-rust	Septoria	Leaf-rust	Yellow rust
	T-positive	0.999	0.997	0.987	0.996
	Precision	0.991	0.998	0.996	0.988
	Recall	1.000	0.997	0.987	0.996
	F-measure	0.995	0.997	0.991	0.992
PART	T-positive	0.997	0.993	0.991	0.992
	Precision	0.991	1.000	0.998	0.975
	Recall	1.000	0.993	0.991	0.992
	F-measure	0.995	0.997	0.994	0.983
J48	T-positive	0.998	0.997	0.989	0.987
	Precision	0.995	0.992	0.996	0.971
	Recall	0.994	0.997	0.989	0.987
	F-measure	0.995	0.994	0.992	0.979
REPTree	T-positive	0.989	0.984	1.000	0.971
	Precision	0.989	0.987	0.987	0.991
	Recall	0.989	0.984	1.000	0.971
	F-measure	0.989	0.985	0.993	0.981

Table 4-8 True Positive, Precision, Recall and F-Measure of Classifiers with Classes

To sum up, the prediction accuracy of selected classifiers indicated the general classification accuracy of the classifiers algorithms. In addition to above-discussed points the True Positive Rate, False positive Rate, Precision, Recall, and F-measure of each class on each algorithm were selected to evaluate accuracy of classifiers in building the models. Consequently, almost all classifiers were registered nearly the same accuracy scores. But, the True Positive rates of the JRip algorithm also evaluated and compared to other all classifiers over all classes and registered average of 99% for each class and less than 1% False Positive rate for each class.

Further, the algorithm also scored the accuracy of 99.61%/ in classifying each class correctly. While scored 0.39% of the instances to be classified incorrectly. Therefore figure 4.3 illustrated the accuracy of each classifier.

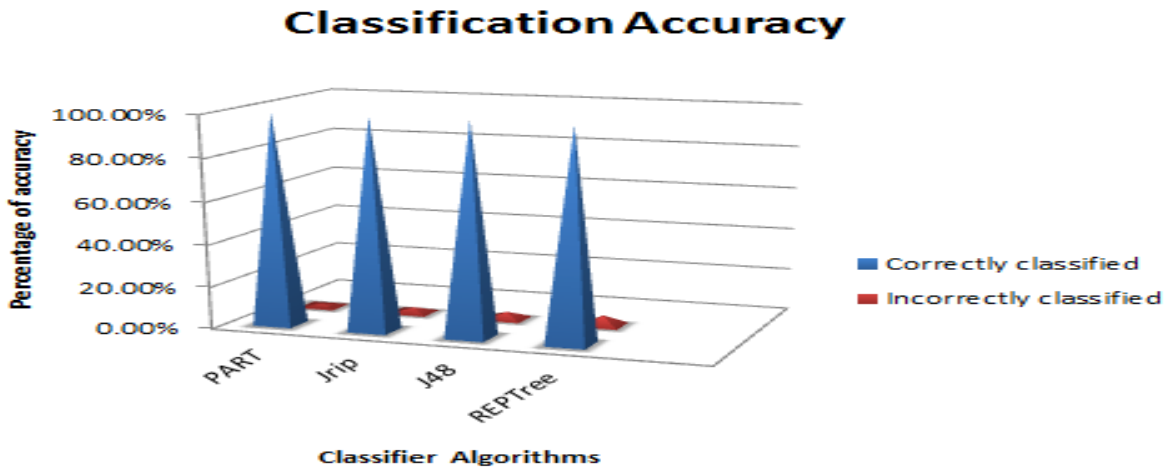

Figure 4-3 Predicted Accuracy of Models by Cross-Validation

Further, the JRip also generated 22 number of rules while classification. Thus depending on objective evaluation of the accuracy of the classifier the JRip performs best and selected for implementing this research work. Figure 4.4 illustrates the time taken to build model for each classifier.

Figure 4-4 Times Taken By Classifiers

The classifiers also evaluated in terms error rates produce while classifying the instances. Accordingly, the JRip and PART classifiers registered minimum percentage in terms of all

measured error rates when compared to other classifiers. However, the REPtree and J48 classifiers registered maximum percentage of errors in order respectively. Here figure 4.5 illustrated the error rates of each classifier.

Figure 4-5 Summarized Error Rates

4.4.3. Evaluating and Using the Rules

Evaluating the rules generated from predictive models set as the main method since it helps to verify whether the rules can be used for the objectives set for. For this study by consulting the domain experts, the rules are evaluated to make sure whether they can identify or diagnosis the Wheat diseases. We performed this by changing rules into natural language, to make understandable for domain experts. Based on the feedback from the domain experts, the rules that have the capability of identifying the diseases were represented in the knowledge base. But the domain experts suggested for further investigation should be done to include all types Wheat diseases with many different attributes especially the symptoms the experts use to identify the diseases in visual practice. Hence the automatic knowledge acquisition task takes into account rules generated from JRip classifier in the integration of data mining induced knowledge with knowledge based system, the study considers the rules generated rom JRip classifier. Table 4.9 shows rules generated by JRip.

JRip generated rules
1. Lesion shape = Elongated-blister AND Favorable conditions = Moderate/Wet: Leaf rust (405.79/1.0)
2. Lesion shape = Elongated-blister AND Parts of crop affected = L-sheath/spike: Septoria (3.61/0.03)
3. Lesion shape = Round-narrow-stripe AND Altitude (m.a.s.l) > 2728 AND District = Hetosa: Yellow rust (160.36/0.36)
4. Degree of damage = Seed-shriveling AND Lesion color = Tan : Septoria (843.75/0.95)
5. Lesion shape = Round-narrow-stripe AND Altitude (m.a.s.l) > 2176 AND Variety = Kubsa AND Altitude (m.a.s.l) <= 2447: Yellow rust (167.91)
6. Lesion shape = Oval-elongated AND Lesion color = Red-brown AND Altitude (m.a.s.l) > 2315 AND Altitude (m.a.s.l) <= 2688 AND Parts of crop affected = Stem/L-sheath/bladder: Stem rust(565.77/0.37)
7. Lesion shape = Round-narrow-stripe AND Altitude (m.a.s.l) <= 2176: Yellow rust (40.93/0.93)
8. Lesion shape = Round-narrow-stripe AND Growth stage = Booting: Yellow rust (30.42)
9. Lesion color = Tan : Septoria (44.49/7.67)
10. Lesion shape = Round-narrow-stripe AND Growth stage = Hard Dough: Yellow rust (16.02)
11. Lesion color = orange-brown AND Parts of crop affected = L-blade, sheath: Leaf rust (38.22/0.63)
12. Lesion shape = Round-narrow-stripe AND Soil Type = Black: Yellow rust (10.0)
13. Lesion shape = Round-narrow-stripe AND Soil Type = Red: Yellow rust (8.0)
14. Degree of damage = Tearing-outer-parts AND Altitude (m.a.s.l) > 2698 AND Lesion color = Red-brown: Stem rust(38.56/2.02)
15. District = Tiyo AND Growth stage = Hard Dough: Stem rust (29.13/0.07)
16. District = Hetosa AND Growth stage = Hard Dough: Stem rust(29.4/0.34)
17. District = Chale: Stem rust (20.74/1.37)
18. District = Hexose: Stem rust (10.12/0.43)
19. Lesion shape = Round-narrow-stripe AND Gene. field condition = Good: Yellow rust (7.34/0.65)
20. Lesion shape = Oval-elongated AND Favorable conditions = Moderate/Wet: Stem rust(16.66/3.88)
21. Lesion shape = Oval-elongated: Septoria (8.1/2.39)
22. : Yellow rust (5.69/0.69)

Table 4-9 Rules Generated From JRip**.**

In order to use the rules obtained from this study the rule-based knowledge representation techniques have been applied, since the result of data mining predictive models and nature of

diseases identification takes the form of rules; hence these rules are easily understandable by human experts. Therefore, the knowledge about Wheat diseases was acquired from the domain experts, in addition to Data mining knowledge discovery methods. The crop diagnosis is the recognizing of its disorder from diseases symptoms and signs. This can be done by identifying whether the disorder is caused by something in the physical environment, an infectious organism (pathogen), and/or an insect. The diagnostic process should include looking at the entire plant as well as its separate parts, carefully analyzing the observations, and attempting to understand or explain why a disorder has occurred. Further, diagnosis of crop disease is the identification of specific crop disease through the symptoms and signs appear on the parts of the crops. In order to begin the treatment process, a disease has to be first diagnosed correctly and completely [59]. Analyzing the results of domain experts' interviews, documents and manuals of Wheat disease identification were undertaken. From the analysis and review result, six main points the pathologists considered when identifying cereal crop diseases were obtained. Since the results obtained in this study also similar to the result obtained in [14], this study also recognized the stated points as it is. Therefore, the analysis made also agreed with the points while acquiring the knowledge from the domain experts and clearly described as the following.

- First, the appearances of crops are observed and compared diseased crop with the normal.
- Parts of crop infected are identified as (head, leaf, stem, root or entire)
- Recognize and understand the characteristics of pathogens or organisms cause infestation.
- Observe symptoms and signs associated with types of disease affecting the Wheat crop.
- Identify and decide the type disease characterized such symptoms and signs.
- Finally, the appropriate treatments for identified Wheat diseases were taken.

Genera, the knowledge applied by the plant pathologists has been identified and documented for each disease. But most of the crop diagnosis activities were performed by observation of the symptoms appeared on the parts of the crop with associated environmental factors. In the rare cases, laboratory examinations were also performed which is the time-consuming procedure. In this research, the symptoms of diseases associated with the types of diseases causing the disorders are identified and represented, depending on the parts of plant affected

as root, stem, leaf and head or grain. The rule-based knowledge representation has been applied for this purpose; hence the rule-based approach is applicable for representing the experience and knowledge of human experts captured in the form of IF-THEN rules and facts. Similarly, it is capable of holding the rules generated from the data mining prediction models. The nature of the disease identification also forced the research to apply rule-based knowledge representation method; because the knowledge acquired from Data mining were rules and the guidelines for diagnosis and treatment of diseases are also of rules. The samples of rules constructed are listed as the following;

Rule1: IF the crop has Lesion color = orange-brown AND

Parts of crop affected are = Leaf bladder &

Leaf sheath AND elongated blister = lesion shape AND Favorable condition = Moderate-wet THEN Leaf Rust.

Rule2: IF crop has Lesion shape = oval Elongated AND

Parts of crop affected = Leaf sheath AND spike AND

Lesion color = Tan AND damage = seed shriveling AND Favorable condition = Cool- wet

THEN Septoria.

Rule3: IF crop has damage = tearing outer parts AND

Has Lesion shape = oval Elongated AND red- brown lesion AND

Altitude > 2315 AND altitude <= 2688 AND Parts affected = Stem, leaf sheath and leaf bladder

THEN Leaf rust.

Rule4: IF crop has Lesion shape = rounded narrow strip AND

Parts of crop affected = Leaf blade and sheath AND sheath AND

Lesion colored = Yellow-orange AND damage = non-tearing AND Altitude > 2728

THEN Yellow Rust.

Rule5: IF crop has Lesion shape = rounded narrow strip AND

Variety = Kubsa AND altitude > 2728 AND Soil type = black and read

THEN Yellow Rust.

4.5. System Design and Architecture

Building the Knowledge Base System by integrating the Data mining prediction results with the Knowledge Base System has been set as an objective of this study. Thus, Data mining predictive models predicted types of Wheat diseases from the data set. Prediction of these diseases undertaken based selected attributes. The JRip rule induction classifier has been selected and applied to the dataset to acquire rules from the data. These are rules represented in the knowledge base; hence corresponding facts added to be evaluated against these rules in Jess library. Since the library contains the rules matching engine called Rete Engine or Rete network. The JRip rules matching with facts can be accessed and fired into user interface or console. The main rules are activated and fired in facts module. Therefore the main goal of integration is to make data mining results easily accessible in Knowledge Base System. The results can be easily accessible for the experts to support them in make better decisions. This can be done by directly calling JRip classifier as WEKA class into Java eclipse environment and making accessible to the interface.

The implementation process contains JRip classifiers that performed better prediction results i.e. JRip rules which are generated automatically from dataset hence it has been selected as best performing classifiers algorithm to implement this prototype KBS. To integrate JRip generated rules into Knowledge Base System, since both WEKA API libraries and JavaFX libraries have been included in Eclipse Java development environment, JRip class can be accessed from WEKA libraries. In this way, the system prototype has been designed and implemented. Running the prototype can be done by loading data set and selecting JRip classifier on it, the rules are directly generated from the dataset and added into JESS to be evaluated and matched against the facts.

Facts are added into working memory by selecting attributes values of the data. The Rete pattern matching engine matches pattern in rules against facts. On the other hand, the PART classifier also integrated into knowledge base; with two options either generating JESS rules with data attributes and by selecting its values, this can be done by saving JESS rules as .clp file extension and/or adding single rules one by one into the knowledge base. Then, Load these rules into prototype to be executed by adding facts. General, to implement results by integrating predictive models generated rules and the Knowledge Base System the implementation architecture framework has been designed. The processes and steps are illustrated as in figure 4.6. As can be observed from the architecture, data mining tools and

techniques can be applied to the data set and resulted in building of predictive models. The models are built based on the selected classifiers algorithms. Thus the models should be evaluated objectively and selected based on models evaluation criteria. The selected model generates rules and these rules have to be validated by domain experts to verify whether they represent features of data or not. Following the validation of rules generated by model; it extracted and encoded into the knowledge base for building Knowledge Base System. The details of integration process and components of architecture have been discussed as the following.

Data Mining Tools and Techniques: the initial steps for building predictive model are selecting and preprocessing the data. According to [66] due to its size and origin, the real world data are prone to missing, noisy and inconsistency which resulted in low mining results. This leads toward need in data preprocessing for filling missing and removing redundant and inconsistent data values. After preprocessing the data at hand different data mining tools and techniques can be applied. For this study, classification technique has been applied. For implementing the classification by building the predictive models WEKA has been applied, hence it is very popular, interactive and contains predefined classifiers algorithms. The tool contains predefined algorithm used for data preprocessing technique and to finalize the classification task.

Model: for building predictive the models that can classify instances into class, different rule induction algorithms can be used. Following the selection of mining tools; the models built on the dataset and evaluated for accuracy and appropriateness. The models selected based on objective evaluation criteria. This resulted in the model with the best accuracy in performing classification and capability to generate rules. For this study top, popular classifier algorithms like JRip, PART, J48, and REPTree has been selected and applied for classification techniques. Consequently, JRip classifier has been selected for rule generation and integrated into Knowledge Base System.

Rule Validation: the rule induction algorithms generate knowledge from the data set in the form of IF-THEN rules. These generated rules should be validated to assure whether its reflection of data set or not. This can be done by consulting domain experts before using rules. The validated rules should be extracted from the knowledge base.

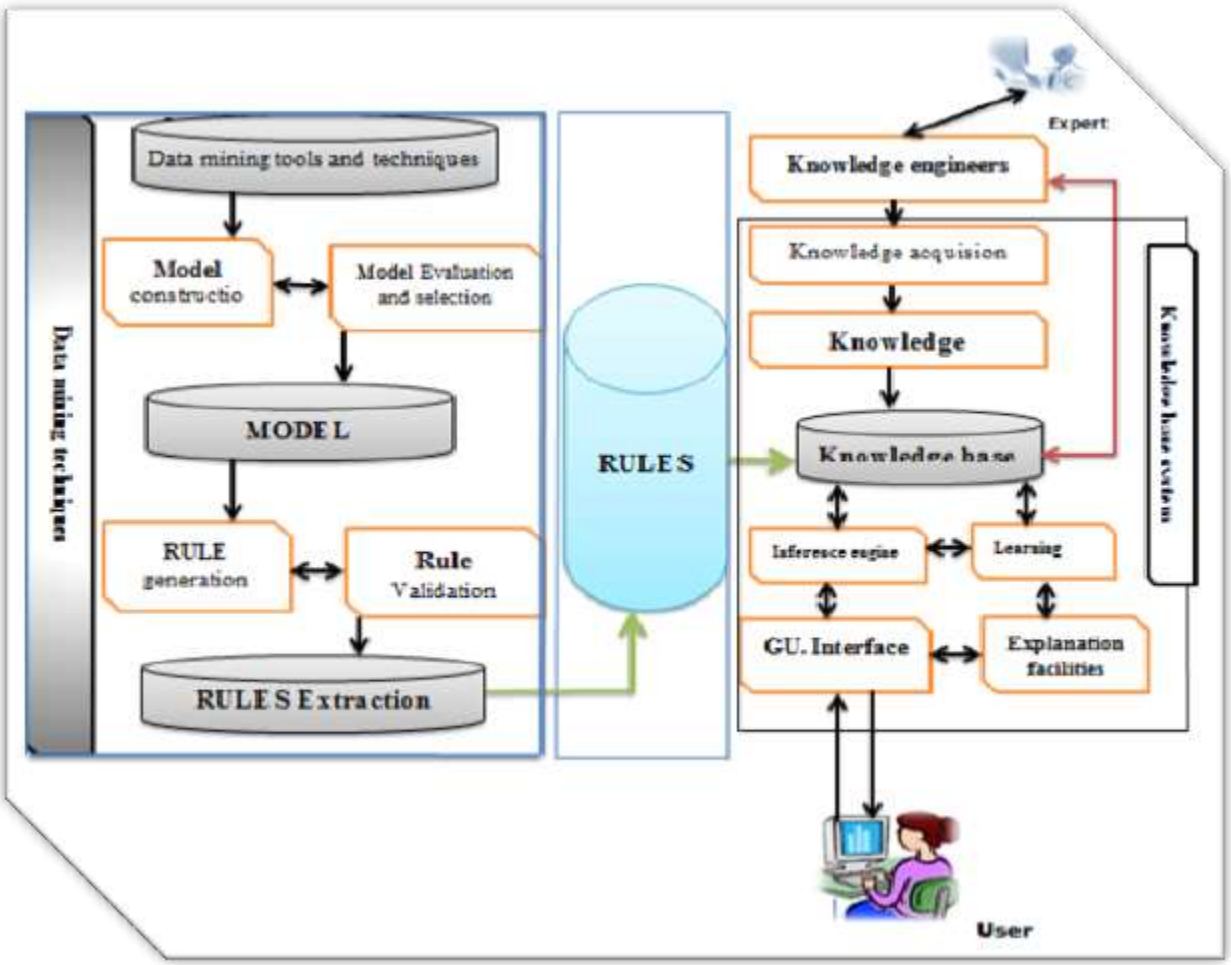

Figure 4-6 Architectural design and Framework for Implementation

Knowledge Acquisition and Representation: The knowledge acquisition is the means by which the knowledge engineer or knowledge base editor acquires knowledge either automatically or manually by interviewing domain experts and document analysis. This extracted knowledge represented in rule-based format and stored into the knowledge base.

Knowledge Base: knowledge base is the repository of knowledge acquired from both rule induction algorithms and manual knowledge acquisition. The knowledge base contains all the knowledge acquired manually and automated data mining knowledge discovery process for detection and treatment of Wheat Diseases. The knowledge stored in the knowledge base can be either of rules and/ or facts.

Inference Engine: An inference engine is the brain of the Knowledge Based System. It directs the system how to derive the conclusion by looking for possible solutions from the knowledge base. It processes all the rules and facts stored in knowledge base to drive the

conclusion. This can be by forward and or backward reasoning mechanisms. For this research forward reasoning mechanism used for inferring solution to the problems. Forward reasoning mechanism is data-driven reasoning approach that starts with facts to reach a conclusion hence JESS rules syntax also support the forward reasoning mechanism.

Extracting Rules the rules generated by rule induction algorithm are extracted into knowledge base directly by calling WEKA classifiers algorithm into Java methods. For this study, JRip classifier has been applied for generating rules. JRip generates rules directly from the data set into prototype's user interface.

Explanation Facilities: is the internal process on how the system reaches its conclusion based on facts and rules stored in the knowledge base. For this study, the implemented prototype provides the conclusion with an explanation. Further, the explanation facility component gives the additional information about detected crop disease and treatment action to be taken.

Learning ability: Is the imitation of human learning; which is an essential component of a knowledge-based system structure. Learning for this architecture is adding new facts into its knowledge base as it runs and remembers these facts at the next time. Thus, knowledge-based system can modify its knowledge as it contains the learning component.

User Interface /GUI; is the interactive area for communicating with the system. The user interface can be graphical user interface (GUI) or command line interface (CLI). The user interface improved to facilitate the communication with users. For this study, Java-based GUI has been developed to access the results in the knowledge base. Thus when the user wants to interact with KBS prototype, has to run the prototype. Since it is public, it does not need any authorization. By selecting the options the user performs any proposed activities like loading data, adding facts and rules and seeing the predicted results.

Knowledge Engineer and Experts: these are external peripherals working in collaboration with the domain experts on how the knowledge acquired, represented and updated if needed in constructed Knowledge Base System.

CHAPTER FIVE

IMPLEMENTATION AND EVALUATION

5.1. Implementations

In the previous chapters, the process to acquire the knowledge from predictive models and how to use the acquired knowledge has been stated clearly. This section mainly focused on describing detailed implementations of obtained results. The most important and interesting objective is integrating the data mining prediction results to Knowledge Base System. Hence the knowledge based system has been constructed to generate the knowledge in the form of rules and facts from the data set and represent in a knowledge base. Therefore the findings of the study have been implemented to make the discovered knowledge accessible to the end users.

To implement the result JRip classifier generated rules automatically from the data and can be accessed from WEKA API directed into Java methods. For this purpose, the following libraries have been used to integrate these results into Knowledge Based System. Java JDK/Java development toolkit/ version8.1 for Java development and run environment of WEKA, Eclipse plugin environment for design and integration, JESS library for Java rule generation and process, JavaFX library and JavaFX scene builder for designing Knowledge Base System prototype. Thus the system tries to identify the diseases into labeled classes and provides description and treatment advice for the detected diseases. Once the system predicts the disease based on selected attributes, it provides description and advice about its treatment for supporting easy of decision making. The system has also capabilities of updating rules and facts as changes happened in data set.

To implement this, the architectural framework has been presented. The knowledge acquisition and representation were undertaken and results were represented as rules, which can be extracted in Knowledge Base directly. Consequently, the knowledge base contains validated rules and facts. In this way inference engine derives conclusions by forward chaining mechanism, starting from the facts and data to get the conclusion. The conclusions are the types of diseases predicted by a classifier based on the attributes values selected by the users via the user interface. It also contains the learning capability to update both facts and rules. This can be the result of either change in a number of instances or total change in data set with the same attributes for identifying diseases. Thus, placement of new rules and facts into

knowledge base is learning capability of the prototype. The prototype is named Wheat Disease Classification Knowledge Base System or WDC-KBS. It contains four modules of detection and treatment of the Wheat diseases. These are;

Data Module: this module contains data and classification algorithms. It also contains all activities for operations like; loading preprocessed and cleaned data, selecting classifiers algorithms and generating rules. For this purpose, the research integrated WEKA classifier as weka.jar into Knowledge Base System to generate rules from data. Thus, JRip classifier has been imported since it is best performing classifier. Therefore the module generates rules from data and sends it to the rule module.

Rules: the module contains the rules generated from the dataset. To interact with this module four buttons needed to be clicked. This includes Send to JESS Button; used to load the selected rules into JESS Working Memory; List Of Rules; used to display all list of rules generated or loaded into system, Add-In Bulk button; used to load many rules and add single Rule provides text area for the user to add list of rules one by one keeping JESS syntax order. For this purpose, the research concerned the JRip generated rules. This contains all the rules generated from JRip and sent to JESS rules to be processed and cross-matched against facts. The rules processed either by selecting all or partials of available rules. These rules can be added by selecting all or Add-in Bulk or can be written into the system by knowledge engineers in keeping the JESS rules syntax.

Facts Module: this module constitutes all the rules for classifiers algorithms and facts added from data by users. For this purpose Rete engine in JESS applied to cross match the rules against the facts. Depending on the rules and facts selected and sent into JESS; the rules were activated by cross-matching against each other. The facts can be added by selecting the attribute values from the drop down option of the prototype or written to it by following the syntax of Jess. Based on the selected attribute, suitable JESS rules should be executed. For instance, if lesion color is selected then the JESS rule which contains disease classification by color should be executed. For firing the activated rules, main rules and facts have to be activated as an internal process. This can be seen in the console interface the module.

Results module: This module contains the results of detected disease types with its descriptions. It also contains treatment and recommendation options to be used by experts or users. This gives the user full information and technical support for understanding the diseases and its treatment.

5.2. Diseases Detection and Activity Flow

The main aim of disease diagnosis is to identify the type of diseases. In this study business problems domain has been analyzed to identify the main threatening diseases in the research area. The identified Diseases are Yellow rust, Leaf rusts, Stem rusts and Septoria since they remain threatening to Wheat crop production in the research area and selected to be analyzed for the study. The detections were undertaken based on rules and facts obtained from predictive model and stored in the knowledge base. These rules are generated from selected attributes with their respective values in the dataset. This section mainly focused to describe the process to detect the diseases as the following. To run the prototype the user selects and loads dataset.

Then, the JRip classifier run on it to generate rules, following this attribute values are selected for each attribute on the lists of the dropdown. These selected attribute values sent as facts into JESS shell to cross match the rules with facts and activate the rule to be fired. These generated rules cross-matched and fired to execute the detection. Optionally the user can select the PART rules and execute. For predicting the Wheat diseases either of two steps can be followed. From the list of rules, either all or some of the rules can be selected and loaded into JESS. Then, corresponding facts should be added by selecting the attribute values for correct detection of the diseases. Consequently, the performed detection ends by clicking on the result menu; where the system provides the description and treatment option for detected Wheat diseases.

Since the user interfaces facilitate and simplify the communication between the system and the user, the prototype developed to implement the result is designed in Graphical based user interface. When the user wants to use the proposed system, s/he has to run the prototype. Then after the following options that would appear to the user and the user can select one of them based on what s/he wants to perform using the proposed system. For this purpose, the activity flow diagram has been designed. Figure 5.1 illustrated the functional flow of activities to be performed to run the prototype.

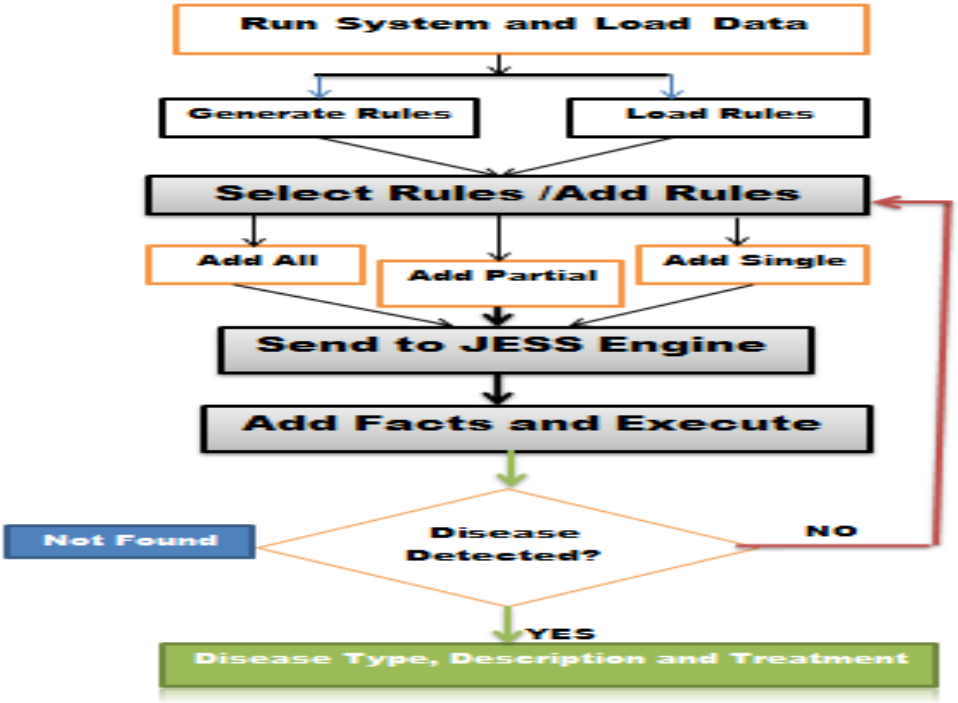

Figure 5-7 Functional Flow of Activities

As it can be observed from the Figure 5.2 the user has four options to interact with it. To enter her/his option from the available four options, the only thing that the user has to do is double-click on the button and then after based on the user chooses the system displays the next user interface.

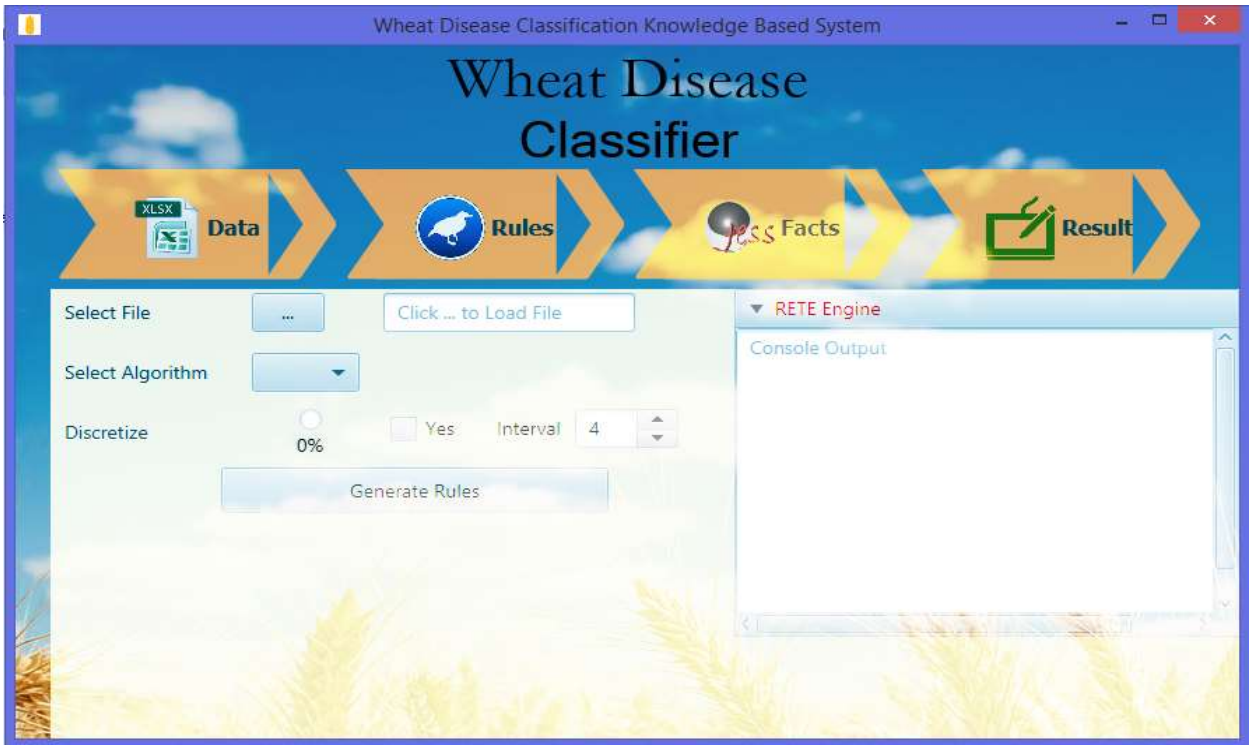

Figure 5-8 Home Page User Interface

For instance, if the user loads the data set the above interface will be displayed as follows: it contains Select File Menu; to select data from specified folder, Select Algorithm Menu; to select classifier algorithm, Discretize Data box; to group numerical data into specified interval, and Generate Rules Menu; to generate rules from dataset. When the user clicks on the generate rules button the next interface will be displayed. This contains the list of rule generated from the dataset. The rules will be opened when the user click List of Rules Menu on the top. The entire rules can be selected and sent to JESS to activate the rules to be fired. Additionally the by clicking Add in Bulk Button many rules can be loaded into the system. Also, the user can write the single rule by clicking Add Single Rule Button.

Figure 5-9 User Interface Showing List of Rules

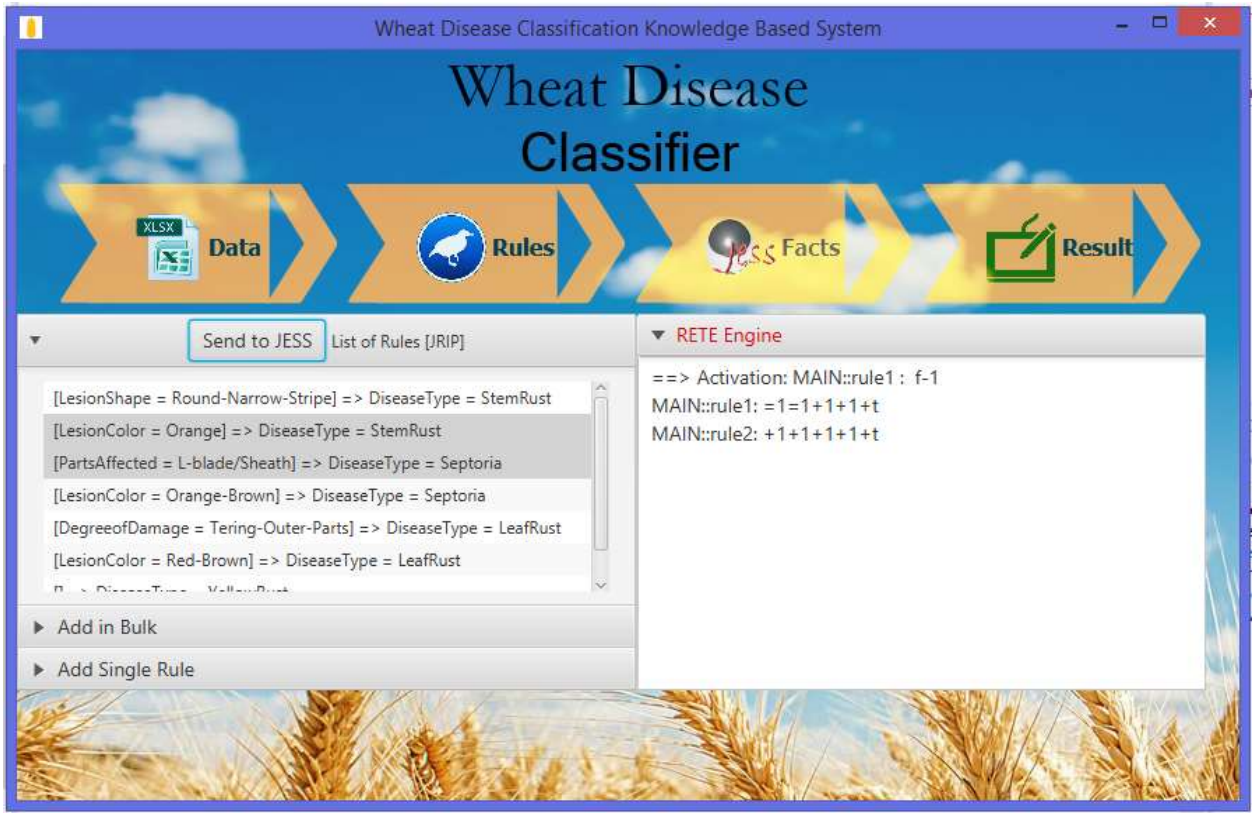

Figure 5-10 User Interface Showing Activated Rules

The list of rules selected sent to the JESS rules will be activated and shown on the console to be fired. Then the corresponding facts have to be added for the activated rules to be fired. Figure 5.3 shows the corresponding interface showing lists of activated rules or the right corner. Before adding the facts the rules have to be added to working memory.

Since the facts are the values of attributes, it can be added by selecting from the top-down selection. By cross-matching the rules in the Knowledge Base with the facts; the inference engine searches the disease with specified symptoms and provides detection results.

Figure 5-11 User Interface for Adding Facts

In the same way, the system will display the following interface for the detected disease. For instance, the Septoria rust disease was detected. The system provides the description and treatment advice for the user.

Figure 5-12 User Interface When Septoria Disease Detected

Figure 5-13 User Interface When Stem rust Disease Detected

The system displays the following interface saying no disease found. The can be the condition when there are no rules matching to the facts added by the user.

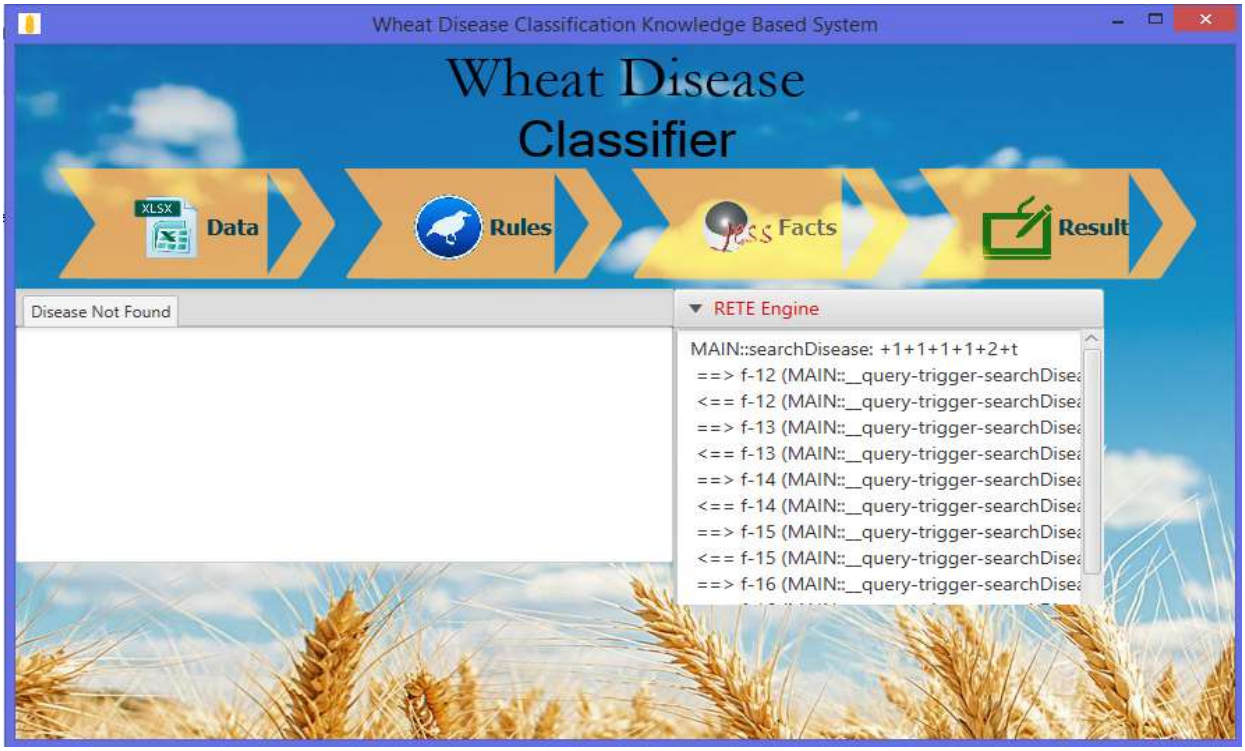

Figure 5-14 User Interface When No Any Disease Detected

5.3. Testing and Evaluation

Testing and evaluation of prototype are the most important steps that are used to measure the performance and user acceptance of the system. Testing is an investigation conducted aiming to assure the system's amenability with its expectation, from the experts and its functioning. It is the measurement that helps to answer questions like: - is the right system? And is the system right? [71]. Accordingly for Wheat disease detection and treatment Knowledge Base System prototype testing and evaluation has been undertaken to implement the result of the study. Thus, testing and evaluation have been undertaken by preparing test cases and providing to the system. Then after, outputs of the system taken to be compared and evaluated with the domain experts' judgments and prototype systems response. Following system's performance testing, the user acceptance testing was undertaken to achieve one of the objectives this study.

5.3.1. System Performance Testing

The System performance testing was undertaken basically to measure how the system performs its purpose accurately. It also performed to assure whether the correct system that meets accuracy level of domain experts has been developed. Additionally, it mainly used to measure whether the prototype performs similar activities with domain expert and operates without the domain experts. For this purpose different system, performance evaluation methods can be applied. Test cases and visual interaction are the commonly used system performance evaluation methods. Since a test case is the most commonly used method, it has been used for evaluating WDC-KBS prototype. Thus, test cases have been prepared to be with domain experts and tested with both domain experts and system. The test case contains 25 instances with 16 attributes per their respective values. Test cases are purposively sampled from dataset with unlabeled instances are given to domain experts.

The same samples were also tested on WDC-KBS. By considering both numbers of attributes taken and times consumed to classify instances into labeled classes, the research compared the output from the system with domain experts' decision. Thus, the table 5.10 illustrates results of evaluation in the form of confusion matric. The columns of the table represent the system's classification, as well as row, represent the domain experts' classification. When analyzed the columns of Stem rust and yellow rust labeled all 5 and 7 instances as Stem rust, yellow rust respectively into their correct classes which indicates both tests scored the same result. Similarly, the column of Septoria also labeled 4 instances into correct class and three

instances incorrectly as yellow rust and it also classifies 4 of Leaf Rust correctly and 2 of them incorrectly classified as Stem Rust.

		WDC-KBS Classification Result				
Domain experts classification	Disease type	Stem rust	Yellow rust	Septoria	Leaf rust	Total
	Stem rust	**5**	0	0	2	7
	Yellow rust	0	**7**	3	0	10
	Septoria	0	0	**4**	0	4
	Leaf rust	0	0	0	**4**	4
	Total	5	7	7	6	**25**

Table 5-10 System Performance Testing Confusion Matrix

In general, the system has classified 20 test case instances out of 25 into their correct classes, which about 80% detection or identification accuracy of the system. Thus, only 5 instances have been classified incorrectly which are about 20% of test case instances. Therefore the experts suggested that it's encouraging performance achievement for the system prototype to be applied in agriculture detecting Wheat disease.

5.3.2. User Acceptance Testing

The user acceptance testing is the activity undertaken to assure how the system performs from the users' perspectives. It also assures the system is accepted and usable for the purpose. Accordingly, the Wheat disease classification Knowledge Base System has been evaluated for assuring its acceptance and applicability in agriculture domain area. For assessing the evaluators feedback different scholars suggested and applied different user acceptance testing standards. Accordingly [72] suggested user acceptance testing standards containing eight points. Similarly, the researchers of [14], [19], [61], [62] and [73] has been adapted and used the same evaluation standards with modification into specific domain area. Thus from the researchers' suggestions and appropriateness of standards, this research work agreed with the scholars' idea. Therefore the standards were customized and used with modification into domain area and used for evaluating this prototype in this research work. Here are lists of the evaluation standards used:-

1. Ease of use and interactive
2. Attractiveness of user interface
3. Efficiency of system in time

4. Accuracy in reaching decision to detect the Wheat disease
5. Incorporation of practical and sufficient knowledge
6. Fitness of system description and treatment of disease
7. Significance of KBS in domain area
8. Table 5-11 User Acceptance Testing and Users Feedback

		Numeric Scales						
No	**Questions**	Poor = 1	Fair =2	Good=3	V/Good=4	Excellent =5	Average	%age
1.	Is the system prototype Easy for you to use and interact with it?	0	0	0	3	2	4.4	88%
2.	Is the WDC-KBS prototype attractive?	0	0	0	4	1	4.2	84%
3.	Is the WDC-KBS prototype more efficient in running time?	0	0	1	1	3	4.4	88%
4.	How do rate the accuracy of the system for reaching decision in detecting the disease?	0	0	0	3	2	4.4	88%
5.	Does the WDC-KBS incorporated practical and sufficient knowledge to detect and treat the Wheat crop disease?	0	0	0	3	2	4.4	88
6.	Does the system fit for describing and treating the Wheat crop disease?	0	1	2	1	1	3.4	82
7.	How do you rate the significance of the system in the domain area?	0	0	0	1	4	4.8	96
average						**Total**	**4.3**	**86.3**

Table 5-12 User Acceptance Testing and Users Feedback

For accessing the evaluator's feedback based on the given standards [18], [73] and [74] recommended numeric weight scales and used for evaluation. Therefore the scale Excellent = 5, Very Good =4, Good =3, Fair =2 and Poor =1. Based on the above evaluation standards the user acceptance testing has been undertaken for WDC-KBS prototype. For this purpose, five domain experts were selected purposively as key informants from Kulumsa agriculture research center to evaluate the system prototype by interacting with it. The evaluators trained on how to run the prototype and told the purpose of evaluation. These selected three researchers and two development agents have been tested the system's performance.

Table Indicates the summary of feedbacks obtained from the evaluators on system's interaction and total evaluation percentage was calculated based on the given scales. Therefore, it summarizes the domain experts' evaluation results. Thus, the value in the table indicates the number of evaluators' responses as excellent, very good, good, fair, and poor, for each respective evaluation standards. Accordingly, for the first evaluation standard that is ease of use and interactive, none of the evaluators respond poor, fair and good. But 60% and 40% of evaluators respond very good and excellent respectively. This resulted in high percentage mainly due to the prototype is designed by the graphical user interface. Secondly, 80% and 20% of the evaluators responded very good and excellent for the attractiveness of the system respectively. Thirdly, 20%, 20%, and 60% of the evaluators responded good, very good and excellent for the efficiency in time of the system respectively. Fourthly, 60% and 40% of the evaluators responded very good and excellent respectively for the accuracy of the in reaching the conclusion. This indicates the majority of the evaluators rate the prototype as very good.

Fifthly, 60% and 40% of the evaluators responded very good and excellent respectively for the inclusion of practical and sufficient knowledge into the system. This indicated the majorities of the evaluators interested in the integration of predictive models result into Knowledge Base System and rate the prototype as very good. Similarly, 20%, 40%, 20% and 20% of the evaluators responded fair, good, very good and excellent respectively for the fitness of the system in describing the detected diseases. Lastly, the regarding the importance KBS in the Agricultural area, 80% and 20% of the respondents responded as excellent and very good respectively. This indicated the knowledge based system's importance to solve agriculture problem. Therefore, the development of integrated Knowledge Base System like WDC-KBS expected to have significant contribution in agriculture, especially in detecting crop diseases. Finally, in summarizing the user acceptance evaluation standards of

respondents the prototype has scored total 86.3%. This can be taken as very encouraging achievement for the prototype to be fully- fledged and implemented in the domain area.

Generally**,** the proposed system prototype achieved promising results of about 80% with the system performance testing and 86.3% user acceptance testing results. This resulted in overall system performance testing of 83.2%. The obtained result became better results when compared to results obtained by [14] in the same domain area; which was achieved 80.9% overall performance by using only interviewing and document analysis. Similarly, it also performs better results than [18] and [61] which uses the data mining as means of knowledge acquisition technique and achieved 80.5% and 81.2% overall performance respectively. In turn, this indicates using integrated (manual and automated) knowledge acquisition techniques results in better knowledge than using either manual or automated knowledge acquisition techniques separately in constructing the Knowledge Based System. The WDC-KBS prototype performs better overall performance results when compared to other research results. Therefore, the prototype has achieved encouraging evaluation results both in terms of user acceptance testing and performance testing for detecting Wheat diseases. As a result, the system registered overall performance of 83.2% which is encouraging result for the system to be implemented in the domain area minimize agriculture problem by supporting the decision making. Consequently, it's promising result to be applied in the agriculture domain area.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6. 1. Conclusion

This research explored and realized the applicability of data mining models and knowledge based system in agriculture for detecting wheat diseases. To realizing the stated objectives, the wheat datasets were collected from Kulumsa Agriculture Research Center and used for building predictive models. For this purpose four experiments were conducted to build predictive models; this mainly aimed at identifying most appropriate and best performing classification model on data. Accordingly, by using objective evaluation standards of models, JRip classification algorithm has been selected since it registered 99.6% overall classification accuracy. This in turn indicated the model has registered promising result for applying data mining models in agriculture for wheat disease predictions. As a result, data mining predictive models can be applied in agriculture for identifying the pattern that determines the wheat diseases. For realizing this, the study used the integrated rule generation knowledge based system. The rules generated from data have been used as patterns to detect wheat diseases. As a result, factors that are determinants for the Wheat diseases occurrences and distribution such as, the date, year, altitude, previous crops, variety, growth stage, field condition, soil types, drainage condition, parts of plant, lesion color and shape, favorable air conditions and degree damage on crop plants are found to be the important and influential for generating patterns relevant for diseases identification. Additionally, the proposed system prototype testing and evaluations were undertaken and achieved promising results with 80% system performance testing and 86.3% user acceptance testing which resulted in overall performance of 83.2%. These are the better results when compared to [14] which was achieved 80.9% overall performance and [18] which achieved 80.5% overall performance by using only data mining as means of knowledge acquisition technique for building knowledge based system. Consequently, the WDC-KBS registered encouraging and acceptable overall performance achievement of 83.2% which is encouraging result to implement integrated predictive models and Knowledge Base Systems in agriculture for crop diseases detection and treatment. Generally, the study achieved better results for minimizing agriculture problems. Therefore, it can be applied in the agriculture domain area.

6.2. Recommendations

The study achieved its objectives by constructing a predictive model to extract hidden knowledge and integrate with Knowledge Base System to make the discovered knowledge accessible for timely routine decision making. Therefore, the study resulted in promising achievement in integrating machine learning-induced patterns with knowledge based system for detecting wheat diseases and providing treatment advice for researchers and development agents.

This study resulted in the promising achievement for further research works to fully implement the integrated machine learning and knowledge based system in the agricultural domain area. As a result, further researches have been suggested as future works based on the observed opportunities and uncovered areas to bring the advances in benefits of integrated data mining results and knowledge based system and forwarded the following recommendations and future works for further study.

- ✓ From the result of this study, the rule based knowledge based system prototype has been developed, tested and found to be applicable in agriculture. It mainly focused on rust disease invasion at research area since other diseases are not registered in a well-organized way. Thus, it recommend that this rule based system can be extended to incorporate all other types wheat diseases such as Fusarium head blight (Calonectria nivalis), Common bunt, pest-related diseases, nutrient deficiency problems and others related problems real Agriculture to detect and provide treatment advice for development agents and researchers to take timely and appropriate actions to overcome problems of outbreaks of rusts and other diseases
- ✓ Since the attained result of this study is promising, Applying integrated machine learning and knowledge based system can be in domain areas other than disease detection, especially to support farmers for advising on farming technologies and farm management.
- ✓ It also recommended the agriculture researchers to apply integrated data mining techniques and approaches other than conventional data analyzing techniques hence the techniques used and obtained results improved the method. Therefore, the methods be extended into other application domain.

REFERENCES

[1] Reta Hailu Belda, "Population dynamism and agrarian transformation in Ethiopia," African Journal of Agricultural Research, vol. 1, sep 2016.

[2] Chen X. M, "Strategies to reduce the emerging Wheat stripe rust disease," ICARDA, International Wheat Stripe Rust Symposium, Aleppo, 2011.

[3] Central statistics agency website available at (2013/2014) "Agricultural sample survey report on farm management practice", accessed February 5, 2016

[4] Bekele Hundie, Verkuiji, H., Mawangi, W. and Tanner, D.G. 2000. "Adaptation of improved wheat technologies in Adaba and Dodola Woredas of the Bale highlands of Ethiopia". CIMMYT/EARO, Addis Ababa, Ethiopia. 2000.

[5] Mulat Demeke "Agricultural Technology, Economic Viability and poverty alleviation in Ethiopia", Addis Ababa University, 1999.

[6] FAO, "wold food and agricultural report on an outbreak of rust " www.fao.org/emergencies/fao-in-action/stories/stories-detail/en/c/451063/ in 2014 /accessed on 25/11/2016.

[7] Tamene Mideksa Serbesa, "Analysis Of Climate Variability On Bread Wheat (Triticum Aestivum L.) Stem rust(Puccinia Graminis F.Sp. Tritici) Epidemics In Bale And Arsi Zones Of Oromia Regional State, Ethiopia," Oromiya, Master Thesis ed. 2015.\7

[8] CALU Factsheet (2009). "Major Pests, Diseases and Weeds of Cereal Crops": available at http://www.pesticides.gov.uk/ accessed on Februry, 23/2016.

[9] Mihaela Oprea, "On the Use of Data Mining Techniques in Knowledge-Based Systems," Economy Informatics, vol. 6, pp. 21-24, April 2006.

[10] Jiawei Han and Micheline Kamber, Data mining: concepts and techniques, 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.

[11] Pedro Domingos, "Toward Knowledge-Rich Data Mining, Data Mining, and Knowledge Discovery," Springer, vol. 15, no. 1, pp. 21-28 , April 2007.

[12] SEO & PPC Management Solution,"December) SEO & PPC Management Solution"[Online].http://www.theecommercesolution.com/usefull_links/KBS_Data_Mining .php, accessed on Februry, 23/2016.

[13] Singh M.P, and Singh, S. (Singh, "Decision Support System for Farm Management World Academy of Science," Engineering and Technology, 2008.

[14] Ejigu Tefera, "Developing Knowledge Based System for Cereal Crop Diagnosis And Treatment: The Case of Kulumsa Agriculture Research Center," M.A. Thesis ed. Addis Ababa University, 2012.

[15] Kassa Belay and Degnet Abebaw, "Challenges Facing Agricultural Extension Agents," Case Study from South-western Ethiopia. Blackwell Publishing Ltd.UK, 2004

[16] Prasad Babu V. A G., "A Study on Various Expert Systems in Agriculture," Georgian Electronic Scientific Journal of Computer Science and Telecommunications., vol. no 4. pp.11, 2006.

[17] Fayyad, S.S. Kashkash K.A. and, Abu-Naser M, "Developing an Expert System for Plant Disease Diagnosis," Journal of Artificial Intelligence, vol. 1, pp. 78-85, 2008.

[18] Abdulkarim Mohammed, "Towards Integrating Data Mining With Knowledge Based System: The Case of Network Intrusion Detection," M.A. Thesis ed. Addis Ababa University, 2013.

[19] Tesfamariam Mulugeta, " Integrating Data Mining Results with the Knowledge Based System for Diagnosis and Treatment of Visceral Leishmaniasis," International Journal Of Advanced Research In Computer And Software Engineering, vol. 5, no. 5, may 2015.

[20] Jay R. Aroson and Ting-Peng Liang Efraim Turban, "Decision Support Systems and Intelligent Systems, 7th ed. New Delhi, India: Prentice Hall of India, 2007.

[21] M. Mehdi Owrang O, "Database systems techniques and tools in automatic knowledge acquisition for rule-based expert system," in *Knowledge-Based Systems Vol 1*. Washington, United States of America:Acadamic Press, 2008, ch. 8, pp. 201-248.

[22] "Agricultural Development Program/Wheat-3/Market & Agribusiness Development," EARI, Addis Ababa, Available at www.ethioagp.org/Ethiopia 2017, accessed on 24/4/2017.

[23] Hailu Beyene, Steven Franzel, and Wilfred Mwangi, "Constraints to increase Wheat production in Ethiopia's small holder sector Tanner," in Sixth Regional Wheat workshop for Eastern, Central and southern Africa, Mexico, CHIMYT,1990

[24] Joyce Jackson, "DATA MINING: A CONCEPTUAL OVERVIEW," Communications of the Association for Information Systems, Volume no. 8, ISSN: 1529-3181, pp. 267-296, 2002.

[25] Kohavi R. and John G., "Wrappers for feature subset selection", Artificial Intelligence, PP. 273–324, 1997

[26] Gilman, D. M. (2000). Data Mining Overview.

[27] Oded Maimon • Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2nd ed. Ramat Aviv, Israel: Springer New York Dordrecht Heidelberg London, 2010

[28] Fayyad R. Uthurusamy and U.M., G. Piatetsky-Shapiro, P. Smyth, 'Advances in Knowledge Discovery and Data Mining'. Massachusetts: (AKDDM), AAAI/MIT Press, 1996

[29] Cios K., Witold Pedrycz, Roman W. Swiniarski, and Lukasz A. Kurgan Krzysztof J. "Data Mining: A Knowledge Discovery Approach ," 2008.

[30] Connolly.T, Begg, C. and Strachan A., "Database Systems: A Practical Approach to Design, Implementation and Management", Second Edition, Ed.: Addison-Wesley New York, 1999.

[31] Dunham H., "Data mining introductory and advanced topics", Upper Saddle River, NJ: Pearson Education, Inc, 2003

[32] Larose T. Daniel, "Discovering Knowledge in "Data, An Introduction to Data Mining" ," ohn Wiley & Sons, Inc., Hoboken, New Jersey, 2005.

[33] Kantardzic M, Data Mining: Concepts, Models, Methods, And Algorithms, Oline Book, Ed.: In W. &. Sons, 2003

[34] Max Bramer, "Principles of Data Mining", Springer, Ed. UK: Portsmouth ,Springer, 2007

[35] Shahrabi and D, Hajizadeh Ardakani. E, ""Application of Data Mining Techniques in Stock Markets:" A Survey," Journal of Economics and International Finance, vol. 2(7), pp. 109-118, 2010.

[36] CALVIN G. MESSERSMITH and JANET D. JINGKAI ZHOU, "Mastering Data Mining: Mining and the Science of Customer Relationship Management," A computer- Based herbicide Injury Diagnostic Expert System," Weed Technology, vol. 19, pp. 486-491, may 2005

[37] Aditi Mahajan and Anita Ganpati, "Performance Evaluation of Rule Based Classification Algorithms," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol.3, no. 10, October 2014.

[38] Adriane B.S. Serapiao and Antonio C. Bann wart, "Knowledge Discovery for Classification of Three-Phase Vertical Flow Patterns of Heavy oil from Pressure Drop and Flow Rate Data," Journal of Petroleum Engineering, vol. 2013, pp. 1-8, August 2010.

[39] Yuan J. , Y., Cui, W. and Zhan, Y. Huang, "Development of a Data Mining Application for Agriculture in IFIP International Federation for Information Processing," Computer and

Computing Technologies in Agriculture, Daoliang Li, vol. 1, no. 258, 2008.

[40] Rahul Ombale, Sagar Ahire, Manoj Dhawade,and Mrs. Prof. R. P. Karande Rajshekhar Borate, "Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in India," International Journal of Novel Research in Computer Science and Software Engineering , vol. 3 , no. Issue 1, pp. (34-37), January-April 2016.

[41] Majumdarn J. Naraseeyappa, S. & Ankalaki, S. Journal of Big Data (2017) 4: 20. https://doi.org/10.1186/s40537-017-0077-4 Publisher Name Springer International Publishing Online ISSN2196-1115

[42] Milovi. C, B. and V. Radojevi C, 2015. Application Of Data Mining In Agriculture, Bulgarian Journal of Agriculture Science, vol.21: pp. 26-34, 2015.

[43] D. Mackworth, A. and Goebel, R Poole, "Computational Intelligence," Oxford University Press,USA, vol. 1, 1999.

[44] Priti Sajja Rajendra Akerkar, "Knowledge-Based Systems", Ed.: http://books.google.com.et/books?id=mQZnd4zmZsoC&printsec=frontcover#v= , 2017 online book. accessed on may 18/2017.

[45] R. A. Sajja, "Knowledge Based Systems" ones and Bartlett Publishers, 2010.

[46] Sajja & Akerkar, "Advanced Knowledge Based Systems," Model, Applications & Research Eds, vol. 1 , pp. 1 – 11, 2010.

[47] Tripathi K P, "A Review on Knowledge-based Expert System: Concept and ArchitectureArtificial Intelligence Techniques - Novel Approaches & Practical Applications,2011.

[48] Investopedia, "Knowledge Engineering Definition," available at http://www.investopedia.com/terms/k/knowledge-engineering.asp#ixzz4hPNkvagd accessed on May 2017.

[49] Jay R. Aroson and Ting-Peng Liang Efraim Turban, Decision Support Systems and Intelligent Systems, 7th ed. India, New Delhi: Prentice Hall of India, 2007.

[50] Leondes Cornelius T., "Knowledge Based System Techniques and applications", 1st ed. San Diego, United States of America: Academic Press, 2005

[51] Alechina N. Knowledge Acquisition, Representation and Reasoning. united kingdom, 2012.

[52] Robinson, B. (1996), Expert Systems In Agriculture And Long Term Research, Journal of

Plant Science, **76**: pp. 611- 617

[53] Henok B. & Burge, "A Case Based Reasoning Knowledge Based System For Hypertension," (2011,1998).

[54] Chen S. "Knowledge Representation and Reasoning Methodology based on CBR Algorithm for Modular Fixture Design," Journal of the Chinese Society of Mechanical Engineers, no. 593-60, 2007.

[55] Sonmenz.C., Artificial Intelligence Notes Reasoning Methods, 1988

[56] Hetal Patel and Dharmendra Patel "A Brief survey of Data Mining Techniques Applied to Agricultural Data," International Journal of Computer Applications , vol. 95, No. 9, pp. 0975 – 8887, June 2014.

[57] Haiguang Wang* and Zhanhong Ma, "Prediction of Wheat Stripe Rust Based on Neural Networks," in CCTA 2011, Part II, IFIP AICT IFIP International Federation for Information Processing, pp. 504–515, Beiging, 2012.

[58] YANG Xiaodong et.al, "Study of Spatial-Temporal Spread Model for Wheat stripe Rust In small scale Based On Bayesian Network," Beijing Research Center for Information Technology in Agriculture, Beijing, China, , no. 100097, 2012.

[59] Saad Razzaq, Kashif Irfan, Fahad Maqbool, Ahmad Farid, Inam Illahi, Tauqeer ul amin Fahad Shahbaz Khan, "Dr. Wheat: A Web-based Expert System for Diagnosis of Diseases and Pests in Pakistani Wheat," in Proceedings of the World Congress on Engineering,Vol I WCE 2008, London, U.K, July 2 - 4, 2008.

[60] Feidu Akmel, "Web Based Agronomy Advisory System: the Case of Cereal Crops," Addis Ababa University, Addis Ababa, Ethiopia, Master's thesis/unpublished October, 2014

[61] Abdi Chali, "An Integration of Prediction Model with Knowledge Base System for Motor Insurance Fraud Detection: The Case of Awash Insurance Company S.C," M.A. Thesis ed. Addis Ababa, Addis Ababa University, 2016.

[62] Tesfaye Fufa, "Integrated Data Mining and Knowledge Based System for Prediction And Treatment Advice of Diabetes," M.A. Thesis ed. Adama Science and Technology University, 2015..

[63] Anand T., and Brachman R. "The process of knowledge discovery in databases", PP. 37–57, 1996

[64] Cabena P., Hadjinian P., Stadler R., Verhees J., and Zanasi A., "Discovering Data Mining: From Concepts to Implementation", Prentice Hall, 1998

[65] Ajith Abraham." Rule-based Expert Systems". Oklahoma State University, Stillwater, OK, USA.

[66] Witten I. H. and Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2rd ed. Amsterdam: Morgan Kaufmann.

[67] Bayehe Mulatu and Stefania Grando (2011). Barley Research and Development in Ethiopia. International Center For Agriculture Research in the dry areas

[68] Hailu Gebre-Mariam, (2003). "Wheat Production and Research in Ethiopia". IAR, Addis Ababa, Ethiopia, available at 68//www.eap.gov.et/sites, accessed on February, 2016

[69] Berry and Linoff, (2004). Data Mining Techniques for Marketing, Sales and CRM, 2nd Ed. 3.2

[70] Saar-Tsechansky M. and Provost F. 2007. Handling Missing Values when Applying Classification Models. In Journal of Machine Learning Research, Vol. 8, pp. 1217-1250

[71] Hassan M. Ghaziri Elias M. Awad, Google Books.[Online]. Available athttp://books.google.com.et/books?id=CI63F2n4N7AC&pg=PA264&lpg=PA264&dq=user+acceptance+testing+for+knowledge+based+systems&source=bl&ots=onepage&q=user%20acceptance, accessed on May 26/2016.

[72] Chen, P. a. (2010). "A User-Centric Evaluation Framework of Recommender Systems,". IN Proceedings of the ACM RecSys. Workshop on User-centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI), pp. 157-164, 2010.

[73] Solomon Gebremariam. "A Self-Learning Knowledge Based System for Diagnosis and Treatment of Diabetes". M.Sc. Thesis, Addis Ababa University, January, 2013.

[74] K.J., Pedrycz, W., Swiniarski, R Cios, "Trends in Data Mining and Knowledge Discovery," In Advanced Techniques in Knowledge Discovery and Data Mining, pp. 1–26. London: Springer/ in comparison among KDP models, chapter one,2010.

ANNEXIES

Appendix I

Detailed Description of Wheat Diseases

No	Types disease and pathogen	Symptoms to identify disease	Favorable condition	Fungicide to be sprayed and rate
1.	Leaf /brown rust (Puccinia recondita)	✓ Small, orangish-brown lesions occurs on the leaves ✓ Blister-like lesions are most common on the leaves ✓ Smaller, more round lesions, and cause tearing of the leaf tissue	Mild warm temperature(70s-80Fº)	✓ Triadimefon 0.5-1 kg/ha for yellow and leaf ✓ Fenpropimorph 1 l/ha for yellow and leaf
2.	Stem /black rust (Puccinia graminis f .sp. tritici)	✓ Blister-like lesions on leaves of Wheat ✓ Reddish-brown spores of the fungus ✓ Bright yellow streaking ✓	Moisture for night time dews	✓ Tebuconazole l/ha for yellow and Stem rust ✓ Epoxiconazole 0.75 l/ha for yellow and stem rust
3.	Stripe/yellow rust (Puccinia triiformis f .sp. tritici)	✓ Small, round, blister-like lesions that merge to form stripes ✓ Yellow-orange color of lesions on the	High moisture	✓ Propiconazole 0.5 l/ha ✓ Triadimefon20%EC 0.65 l/ha ✓ Tebuconazole 0.65 l/ha
4.	Septoria tritici blotch (Septoria tritici)	✓ Tan, elongated lesions on Wheat leaves ✓ Lesions may have a yellow margin ✓ Darker color leaf blotch or tan spot	High moisture and high humidity	✓ Propiconazole 0.5-1 l/ha ✓ Flutriafol 0.5-1 l/ha ✓ Chlorothalonil 1.125 kg/ha for septoria blotch only ✓

5.	Common bunt (Tilletia caries)	✓ kernels have gray green color and are wider than healthy kernels ✓ Seeds are easily broken during grain harvest ✓ Releasing masses of black powdery spores ✓ Spores produce fishy odor	Cool temperature	
6.	Fusarium head blight (Calonectria nivalis)	➢ Dark brown discoloration at the base of portion of the glume ➢ Wheat grain has White chalky appearance ➢ Tan or light brown lesions on Wheat heads ➢ Spores are released from the head	Wet and humid weather	✓ Chlorothalonil 1.125 kg/ha for septoria blotch only
7.	Barley yellow dwarf (BYD luteovirus)	➢ A flame-like appearance of Wheat leaves ➢ Shorter length of the Wheat crop than neighboring healthy Wheat ➢ Wheat leaves to have a yellow or red discoloration	Vectored by an aphid	✓ Triadimefon 0.5-1 kg/ha for yellow and leaf ✓ c
8.	Powdery mildew (Erysiphe grarninis f. sp. tritici)	➢ White, cottony growth of the fungus ➢ White lesions on leaves and leaf sheaths.	Cool (15-22 ^{0}C), Cloudy, and humid (75-100% relative humidity)weather	✓ Triadimefon 0.5-1 kg/ha for yellow and leaf
9.	Common root rot (Helminthosporium)	➢ Patches of white heads scattered throughout a field ➢ Premature death of Wheat	Crop Residue in the farm	✓ Chlorothalonil 1.125 kg/ha for septoria blotch only ✓

10.	Barley yellow dwarf (BYD luteovirus)	✚ A flame-like appearance of Wheat leaves ✚ Shorter length of the Wheat crop than neighboring healthy Wheat ✚ Wheat leaves to have a yellow or red discoloration	Vectored by an aphid	✓ Triadimefon 0.5-1 kg/ha for yellow and leaf
11.	Powdery mildew (Erysiphe grarninis f. sp. tritici)	➢ White, cottony growth of the fungus ➢ White lesions on leaves and leaf sheaths.	Cool (15-22 ^{0}C), Cloudy, and humid (75-100% relative humidity)weather	
12.	Common root rot (Helminthosporium)	➢ Patches of white heads scattered throughout a field ➢ Premature death of Wheat	Crop Residue in the farm	✓ Chlorothalonil 1.125 kg/ha for septoria blotch only
13.	Take-all (Gaeumannomyces graminis f, sp. tritici)	✓ Black discoloration of the lower stem and roots ✓ Wheat to die prematurely	Cool Soil T^0 and nutrient deficient soils	
14.	Bacterial streak (Bacillus megaterium pv. cerealis	✓ Water-soaked areas become tan streaks within a few days ✓ Lesions dry to form a clear, thin film on leaf ✓ Water soaked areas between leaf veins	Wheat curl mite	✓ Triadimefon 0.5-1 kg/ha for yellow and leaf

Appendix II

Lists and Descriptions of all Attributes

No.	Name of Attributes	Data Types	Description
1.	Year	Numeric	Describes the year at which the data has been collected. Contains 3 distinct numeric values i.e. 2014,2015,and 2016
2.	Date	Nominal	Describes the Specific date at which the disease survey was taken from fields. Contains five distinct nominal values in months: i.e. August, September, October, November, and December,
3.	District	Nominal	Describes the specific research area from where the data were collected. Contains 22 nominal distinct values i.e. G/Asasa, Nagele, shaashemene, Kofale, Dodola, Adaba, Tiyo, Hetosa, Lodehetosa, Diksis, Arsi robe, Ticho, Kula/Sude, Sire, Chale, Guna, Marti, Gonde, Itaya, hunkolowabe, Lemu-bilbilo, Digalu,
4.	Altitude (m.a.s.l)	Numeric	Describes the measure height of place above the sea level. Contains 115 numeric distinct values.
5.	Previous crop	Nominal	Describes the crops which have been sawed on the fields before the season the survey was taken. Contains 7 nominal distinct values i.e. Wheat, maize, barley, potato, fallow, fababean, and maize and teff…
6.	Variety	Nominal	Describes the types Wheat crops adapted in the area. Contain 18 nominal distinct values i.e. Kakaba, Kubsa, Danda, Digalu, kubsa+mixture, Paveen, digalu+mixture, K62954A, Triticale, Israel, unknown, ET13A2, maddawolabu, Hulluka, Hidase,Ogolcho, enkoy, and shorima …
7.	Growth stage	Nominal	Describes the series of steps passed by the crop beginning from saw to harvest. Contains 14 nominal distinct values i.e. Dough, Early-dough, Soft-dough, Hard-dough, heading, Early-flowering, Flowering, flowering-complete, booting, Milky, Milky–soft-dough, milky-flowering, Tillering complete and maturity.
8.	Field condition	Nominal	Describes the general condition of the field from which the survey was taken. Contains 3 nominal distinct values i.e. good, moderate and bad.
9.	Soil Type	Nominal	Describes the Types of soil from the fields of surveyed area. Contains 5 nominal distinct values i.e. light(sand), Black(heavy), Red(clay), Red and black.

10.	Drainage Condition	Nominal	Describes the condition at water is able flow in soil. Contains 3 nominal distinct values i.e. good, bad and moderate.

11.	Favorable conditions	Nominal	Describes the condition of temperature which is favorable for development of Wheat rust. Contains 3 nominal distinct values i.e. cool/wet, low/wet, moderate/wet
12.	Parts of crop affected	Nominal	Describes the parts of crop affected by the diseases. Contains 3 nominal distinct values i.e. stem/leaf-sheath, leaf-blade, leaf-sheath/spike.
13.	Degree of damage	Nominal	Describes the degree of damage seen on parts affected of the crop. Contains 3 nominal distinct values i.e. tearing outer parts, seed-shriveling, non-tearing.
14.	Lesion color	Nominal	Describes the colors of lesions observed on the affected parts of crops. Contains 4 nominal distinct values i.e. reddish-brown, orange-brown, tan and orange.
15.	Lesion shape	Nominal	Describes the shapes of lesions observed on the affected parts of crops. Contains 3 nominal distinct values i.e. oval- elongated, elongated-blistered and round-narrow-stripe.
16.	Disease type	Nominal	Describes the Disease types are the divisions of the Wheat crop rusts. Contains 4 nominal distinct values i.e. yellow rust, leaf rust, Stem rustand septoria.
17.	Season type	Nominal	Describes the Specific time season at which the production was undertaken. Contains single nominal distict value i.e. main season.
18.	Region	Nominal	Describes the specific area at which the research area where found. Contains single nominal distinct value i.e. Oromiya.
19.	Zone	Nominal	Describes the specific research area from where the data were collected and research conducted. Contains single nominal distinct value i.e. Arsi
20.	Type of wheat	Nominal	Describes the types Wheat crops adapted in the area and mostly used for study. Contain single nominal distinct value i.e. bread wheat.

Appendix III

Interview questions

The following structured interview questions are prepared for domain Experts and researchers focusing on Wheat crops diseases identification and treatment processes, at Arsi. The main objective of conducting the interview is to investigate the construction of a predictive model and integrate with Knowledge Base System for detecting and treating Wheat disease that could help in the effort of enhance and improve the accuracy of decision made experts and development agents.

1. What are the challenges that face the research center to achieve its objectives?
2. Is there any factor that inhibits the research center in innovating, disseminating and transfer of knowledge and research outputs to stakeholders?
3. If your answer in Question No 1, is yes, which of the following is inhibiting factors?
   a. Lack of experienced experts and high turnover of experts at the research center.
   b. Lack of infrastructure and sophisticated analysis and interpretation tools like data mining, knowledge based, decision support systems and professionals in this at the research center.
   c. Please specify others, if any_____?

Questions concerning Wheat crop diseases

4. What are the types of diseases that affect Wheat particularly in Ethiopia?
5. What are the symptoms that observed for the diseases identified in Question No.4 above?
6. What are the identification techniques and procedures are applied to diagnose these diseases?
7. Do you use standardized guidelines/manuals to diagnose each crop diseases?
8. What are the steps undertaken to diagnose the each diseases each mentioned above?
9. After diagnosis of these diseases, what are the recommendation and treatments you provide?

Appendix IV

User Acceptance Test

Dear Evaluator,

This evaluation form is prepared to measure acceptability and usability of WDC-KBS prototype by the users in the area of agriculture. Further, it also measures the applicability the prototype knowledge based system to detect and treat Wheat diseases in agricultural area. Therefore, you are kindly requested to evaluate the system's prototype by labeling (√**)** symbol in the boxes 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 scales in the table are rated as numeric values of: -

No.	Scale	Numeric values	Please put total here
1.	Excellent	5	
2.	Very good	4	
3.	Good	3	
4.	Fair	2	
5.	Poor	1	

Table of the evaluation standards' numeric values

1. Is the system prototype Easy for you to use and interact with it?

   A. Poor ☐ B. Fair ☐ C. Good ☐ y good . ☐ lent ☐

2. Is the WDC-KBS prototype attractive?

   A. Poor ☐ B. Fair ☐ C. Good D ☐ good E. ☐ llent ☐

3. Is the WDC-KBS prototype more efficient in running time?

   A. Poor ☐ B. Fair ☐ Good D ☐ good ☐ llent ☐

4. How do rate the accuracy of the system for reaching decision in detecting the disease?

   A. Poor ☐ B. Fair ☐ C. Good ☐ y good E ☐ ellent ☐

5. Does the system fit for describing and treating the Wheat crop disease?

   A. Poor ☐ B. Fair ☐ C. Good D ☐ good ☐ llent ☐

6. How do you rate the significance of the system in the domain area?

   A. Poor ☐ B. Fair ☐ C. Good D ☐ good ☐ llent ☐

7. How do you rate the significance of the system in the domain area?

   B. Poor ☐ B. Fair ☐ C. Good ☐ very good ☐ Excellent ☐

Appendix V

Sample codes

```
package application;

import java.io.File;

import java.net.URL;;

import javax.swing.event.DocumentListener;

import javafx.stage.FileChooser;

import jess.JessException;

import jess.Rete;

import jess.swing.JTextAreaWriter;

import weka.classifiers.rules.JRip.Antd;

import weka.classifiers.rules.JRip.RipperRule;

public class WekaRulesController extends
VBox implements Initializable {

	Rete engine =
Begin.getReteInstance();

	@FXML

	private Text txtStat1;

	@FXML

	private Text txtStat2;

	@FXML

	private Text txtStat3;

	@FXML

	private Text txtStat4;

	public static ArrayList<Integer>
selected = new ArrayList<Integer>();

	@FXML

	private ListView<String> lvRules;

	@FXML

	private TextArea txtNewRule;

	private TextArea txtRConsole;

	private JTextArea txtConsole = new
JTextArea();

	JTextAreaWriter taw = new
JTextAreaWriter(txtConsole);

	@FXML

	private Button btnSend;

	@FXML

	private TextField bulkURL;

	private File file;

	public ListView<String> getLvRules()
{

		return lvRules;

	}

	public void
setLvRules(ListView<String> lvRules) {

		this.lvRules = lvRules;

	}

	public void initialize(URL arg0,
ResourceBundle arg1) {

	txtConsole.getDocument().addDocum
entListener(new DocumentListener(){

			@Override

			public void
changedUpdate(DocumentEvent arg0) {


	txtRConsole.setText(txtConsole.getTe
xt());

			}

			@Override

			public void
insertUpdate(DocumentEvent arg0) {

	txtRConsole.setText(txtConsole.getTe
xt());
```

```
                        }

                        @Override

                        public void
removeUpdate(DocumentEvent arg0) {
        txtRConsole.setText(txtConsole.getTe
xt());

                        }} );

                try {

                engine.addOutputRouter("t",
taw);

        engine.addOutputRouter("WSTDOUT
", taw);

        engine.addOutputRouter("WSTERR",
taw);

                        engine.run();

                        engine.watchAll();

                } catch (JessException e) {

                        e.printStackTrace();}


        lvRules.getSelectionModel().setSelect
ionMode(SelectionMode.MULTIPLE);


        lvRules.getItems().removeAll(lvRules.
getItems());

                if (DataController.getRules()
!= null && DataController.getRules().size() >
0) {

                        for (RipperRule r :
DataController.getRules()) {

lvRules.getItems().add(r.getAntds().toString()
+ " => " +
DataController.data.classAttribute().name() +
" = " +
DataController.data.classAttribute().value((int)
r.getConsequent()));

                        }                   } else
{


        System.out.println("JRIP rules not
found"); } }

        public void sendRules(ActionEvent
event) throws Exception {

                selected.clear();

                Rete engine =
Begin.getReteInstance();

                engine.watchAll();

                Iterator<Integer> iter =
lvRules.getSelectionModel().getSelectedIndic
es().listIterator();

                int index;

                while (iter.hasNext()) {

                        index =
Integer.parseInt(iter.next().toString());

                        selected.add(index); }

                if (DataController.getRules()
!= null && DataController.getRules().size() >
0) {

                        int ruleCounter = 0;

                        for (RipperRule r :
DataController.getRules()) {

                                String
ruleString = "(defrule  rule" + ruleCounter + "
";

                                if
(r.hasAntds() == true) {


        Iterator<?> antIterator =
r.getAntds().listIterator();

                while (antIterator.hasNext()) {

Antd ant = (Antd) antIterator.next();
                        String name =
ant.getAttr().name(); String value =
DataController.data.attribute(name).value((int)
ant.getAttrValue());
        ruleString += "(" +
name.replaceAll("[^\\w\\s]", "").trim() + " "

+ value.replaceAll("[^\\w\\s]", "").trim() + ") ";

}       }

ruleString += " => ";
```

```
ruleString += "(assert (DiseaseType ";

String trimmed =
DataController.data.classAttribute().value((int)
r.getConsequent())

.replaceAll("[^\\w\\s]", "").trim();

ruleString += trimmed;

ruleString += ")))";

if (selected.contains(ruleCounter)) {

engine.executeCommand(ruleString) }

ruleCounter++;}

}

//txtStat1.setText("Successfully sent");}

public void executeSingleRule(ActionEvent
event) throws Exception {

engine.executeCommand(txtNewRule.getText
());

        }

        public void browseClick(ActionEvent
event) throws Exception {

                try {

                        FileChooser
fileChooser = new FileChooser();


        fileChooser.setTitle("Open JESS
script");

                        file =
fileChooser.showOpenDialog(Begin.theStage)
;


        bulkURL.setText(file.getAbsolutePath
().toString());

                } catch (Exception e) {


        System.out.println("File operation
error");} }

        public void
executeBulkRules(ActionEvent event) throws
Exception {

        engine.executeCommand(txtNewRule.
getText());

        System.out.println(file.toURI().getPat
h());

        engine.batch(file.toURI().getPath());}
}
```