Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Integrating Data Mining Results with the Knowledge Based System for Diagnosis and Treatment of Visceral Leishmaniasis

[1]Tesfamariam Mulugeta Abuhay, [2]Tibebe Beshah Tesema
[1]Department of Information Systems, University of Gondar, Gondar, Ethiopia
[2]School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia

Abstract— the general objective of this study is to investigate the construction of Visceral Leishmaniasis Knowledge Based System through Integrated Knowledge Acquisition Technique. Mixed research design, integrated knowledge acquisition techniques, CommonKADS system development methodology, rule based knowledge representation approach, SWI-Prolog 6.4.0 with UTF-8 and Java NetBeans IDE 7.3 with JDK 6 were used to achieve the objective of the study. Subjective and Objective interestingness measures were used to evaluate the results and accuracy of the data mining model. In addition to this, system performance and users' acceptance testing were used to evaluate the prototype KBS.

To identify the best prediction model for treatment of VL, three experiments for three classification algorithms were conducted. Finally, the researchers decided to use rules of PART classification algorithm for further use in the development of knowledge base of KBS because it registered better performance with 67.5 % and 75.4% of Objective and Subjective evaluation result, respectively. Test cases and questionnaire were prepared to evaluate the performance and users' acceptance of the proposed system. As a result, the proposed system can perform in the absence of domain experts with 95% of system performance evaluation result which indicates that the KBS has the necessary knowledge for diagnosis and treatment of VL which in turn implies that the study was effective in acquiring knowledge. Besides to this, the proposed system achieves 86% of the users' acceptance which in turn implies that the proposed system could be operational if it could be implemented.

Keywords— Leishmaniasis, Knowledge Based System, Data Mining, Knowledge Acquisition and Knowledge Representation Techniques.

I. INTRODUCTION

Visceral Leishmaniasis, or kala azar as it is more commonly known, which affects the immune system is caused by a parasite transmitted by sand flies. The disease is endemic in environments that range from deserts to rain forests in rural and urban settings in over 98 countries of the tropics, subtropics, and southern Europe. Globally, the population at risk is estimated to amount to 350 million people with an overall prevalence of 12 million [1, 3, 6, 8, 11].

In Africa, kala azar is found in parts of Sudan, Kenya, Somalia, Eritrea and Ethiopia. According to the 2012 WHO global Leishmaniasis estimate, Ethiopia is one of the ten high burden countries for VL. It is estimated that the annual burden of VL ranges from 2,000 to 4,500 cases. The disease is particularly prevalent in the lowlands of northern Ethiopia, among the large sesame, cotton and sorghum farming areas. Here, hundreds of thousands of migrant workers arrive every year to work the agricultural season. They are at a high risk of getting kala azar as they live and work in difficult conditions and often lack adequate nutrition, clean water, shelter and protective equipment. In recent years, kala azar has also spread to a number of areas in the highlands of the Amhara region and parts of the Southern Nations and Nationalities People's Region [1, 3, 6, 8, 11].

Even though many steps have been taken in pushing forward the response to kala azar in Ethiopia and many lives saved as a result, the Data Analysis and Visualization result shows that the number of patients is increasing from year to year and most of them are young and productive who could contribute a lot for the development of their country. Hence there remains a great need for increased preventative measures and accessibility to early diagnosis and effective treatment if the disease is ever to be eliminated [3, 11].

Besides to this, since there is also lack of senior educated doctors in the country, one of the steps that Ethiopian Ministry of Health has taken is giving training for health workers and placed them in the rural and remote areas of the country. However, as one of the researcher observed during his time in Abdurafi, these health workers are not capable enough to give appropriate medical health care services and treatment for patients because the knowledge of these persons about how to diagnosis a person with VL and give treatment is not adequate enough because these health works are not well trained and experienced in this specific disease. Due to this, most of the patients prefer to come to Gondar which is a regional town where relatively better service can be accessed and senior health works are available. We also interviewed patients in the Leishmaniasis Research and Treatment Center of the University of Gondar Hospital by asking why they prefer to come to Gondar and they replied that the health care service in places where VL is endemic is not good enough to cure them.

Therefore, in order to help health workers to upgrade their knowledge and experience and give high level medical health care services and treatment, developing a Knowledge Base System that gives advice in Amharic Language, which is a national language of Ethiopia, about VL will play a critical role so that health workers can save many lives. Besides to this, developing the knowledge based system in Amharic language will also encourage and enable patients who have computer access to diagnosis themselves and check the severity of the disease and go to higher clinic for further diagnosis.

Knowledge Base System or Expert system is a computer program that simulates the judgment and behaviour of a human that has expert knowledge and experience in a particular field such as medicine where there is a shortage of expert knowledge. During development of knowledge based system, knowledge must be acquired about the problem to be solved because the most important ingredient in any expert system is the knowledge and the power of expert system resides in the specific, high-quality knowledge it contain about task domains. Knowledge can be acquired from different sources such as making interview with domain experts, document analysis, observation and others. Since tacit knowledge is personal and the knowledge expert may not tell all the knowledge s/he knows during interview, there is hidden knowledge about the problem. To alleviate this problem, automatic knowledge acquisition is proposed. Therefore, Data mining, more general knowledge discovery techniques, proposed for extracting hidden and previously unknown knowledge from datasets by different researchers [2, 5, 12].

II. SIGNIFICANCE

Even though they are short in number, many educated man powers particularly senior doctors and nurses are not willing to go to VL endemic places. Due to this, persons who are infected with VL around VL endemic place are left untreated. In addition, patients could not go to places where VL diagnosis and treatment is available because it costs them huge amount of money. Therefore, developing and implementing Knowledge Base System that provides advice to health workers and patients in Amharic language would play a great role in providing early diagnosis and effective treatment so that the health care service could be improved because KBS provides the high-quality performance which solves difficult problems in a domain as good as or better than human experts and can possesses vast quantities of domain specific knowledge to the minute details [5] which can in turn serve as means of knowledge and experience transfer because the ability of the Knowledge Base System to capture and redistribute expertise has significant implications on development of a nation, commodity or population. In addition to this, such systems allow documentation of one or more expert knowledge and utilize the knowledge for problem solving in cost effective way. It allows for, in a controlled manner, the import of expertise in various areas that the nation lacks, the export of knowledge relating to domestic areas of expertise, and the duplication and redistribution of scarce knowledge in a cost effective manner. Thus areas of expertise that the selected domain/region/nation is deficient in or possesses exclusively are potential candidates of the knowledge-based systems [15].

III. KNOWLEDGE ACQUISITION

The development of an efficient knowledge-based system (KBS) involves the development of an efficient knowledge base that has to be complete, coherent and non redundant. The step of knowledge acquisition is one of the major bottlenecks in the stage of knowledge base development. In order to make knowledge extraction as much as correct as possible (i.e. in order to keep the correctness of the knowledge as it is kept at the source) different techniques such as interview, questionnaires, documents analysis and observation could be applied. In addition to this these techniques, data mining techniques, more general, knowledge discovery techniques became the most used in the recent years [16]

For this study, the researchers acquire knowledge using integrated knowledge acquisition techniques which are documents analysis, interview and data mining.

A. Manual Knowledge Acquisition (Documents Analysis and Interview)

We used four documents which are Guideline for Diagnosis, Treatment and Prevention of Leishmaniasis in Ethiopia which is prepared by Ethiopian Federal Ministry of Health, MSF Holland guideline for diagnosing and treatment of VL, Ministry of Health Uganda guide line for diagnosis and treatment of VL and an article called Visceral Leishmaniasis which is prepared by Johan van Griensven and Ermias Diro.

We also conducted interview with Doctors, Nurses, Laboratory Technicians and Pharmacists that works in the Leishmaniasis Research and Treatment Center of the University of Gondar and we found that the knowledge acquired from documents and interview are complementary and used the result as data triangulation. Accordingly we presented the knowledge that is acquired from documents and interview as follows:

1) *What is Leishmaniasis:* Visceral leishmaniasis (VL), also known as kala-azar, is a disseminated protozoan infection caused by Leishmania donovani complex and is a neglected but typically fatal vector-borne protozoan disease reported from all continents except Antarctica and Australia [2, 6, 8, 11].

In Ethiopia, Visceral Leishmaniasis is found mainly in the lowlands of northwest, central, south and southwestern Ethiopia; whereas cutaneous Leishmaniasis is widely distributed all over the country. Several outbreaks occurred in the last few years. The disease is spreading to new localities, namely, Amhara, Tigray, Southern Nations, Nationalities, and Peoples' Region, Oromia and Somali. In addition, there have been recent outbreaks in northern and southern parts of the country: Libo Kemkem Woreda in Amhara region, T/Adiabo Woreda in Tigray region and Imey in Somali region, due to population movement and HIV co-infection [11].

According to the 2012 WHO global Leishmaniasis estimate, Ethiopia is one of the ten high burden countries for VL. It is estimated that the annual burden of VL ranges from 2,000 to 4,500 cases. Some of the factors found to be associated with the spread include population movements to and from endemic focus areas, poverty and malnutrition associated with presence of the sand fly vector and reservoirs [11].

2) Infection and transmission: Leishmaniasis is transmitted by sand flies. Sand flies cannot tolerate strong heat and light and need a cool and humid resting place during the day. At night, they feed on sugar meals taken from plants, but female sand flies, needing protein for their eggs, also take a nocturnal blood meal and this is when Leishmania infections are acquired and transmitted [2, 6, 11].

There are 2 different transmission cycles for kalaazar (KA): Anthroponotic, with only humans as the source of infection for the vector and Zoonotic, with wild or domesticated animals (dogs) asreservoir hosts [2, 6, 11].

3) *Clinical Diagnosis:* Clinical diagnosis of KA has been demonstrated to be highly unspecific. In East Africa, the prevalence of KA in clinical suspects was between 30-60%.Therefore, clinical suspicion of KA must always be followed by a demonstration of the presence of the parasite either directly (microscopy) or indirectly (serology) [6, 8].

A person who presents with fever for more than two weeks and an enlarged spleen (splenomegaly) and/or enlarged lymph nodes (lymphadenopathy), or either loss of weight, anemia or leucopenia while living in a known VL endemic area or having travelled to an endemic area [6, 8, 11].

4) *Laboratory Diagnosis:* As the clinical presentation of VL lacks specificity a definitive test is required to decide which patient should be treated. This test has to be highly sensitive (> 95%) and specific. Ideal tests should be able to make distinction between acute or asymptomatic infections. Tests available to diagnose VL include non-Leishmanial tests, parasite detection, antibody detection, antigen detection, and molecular techniques [11].

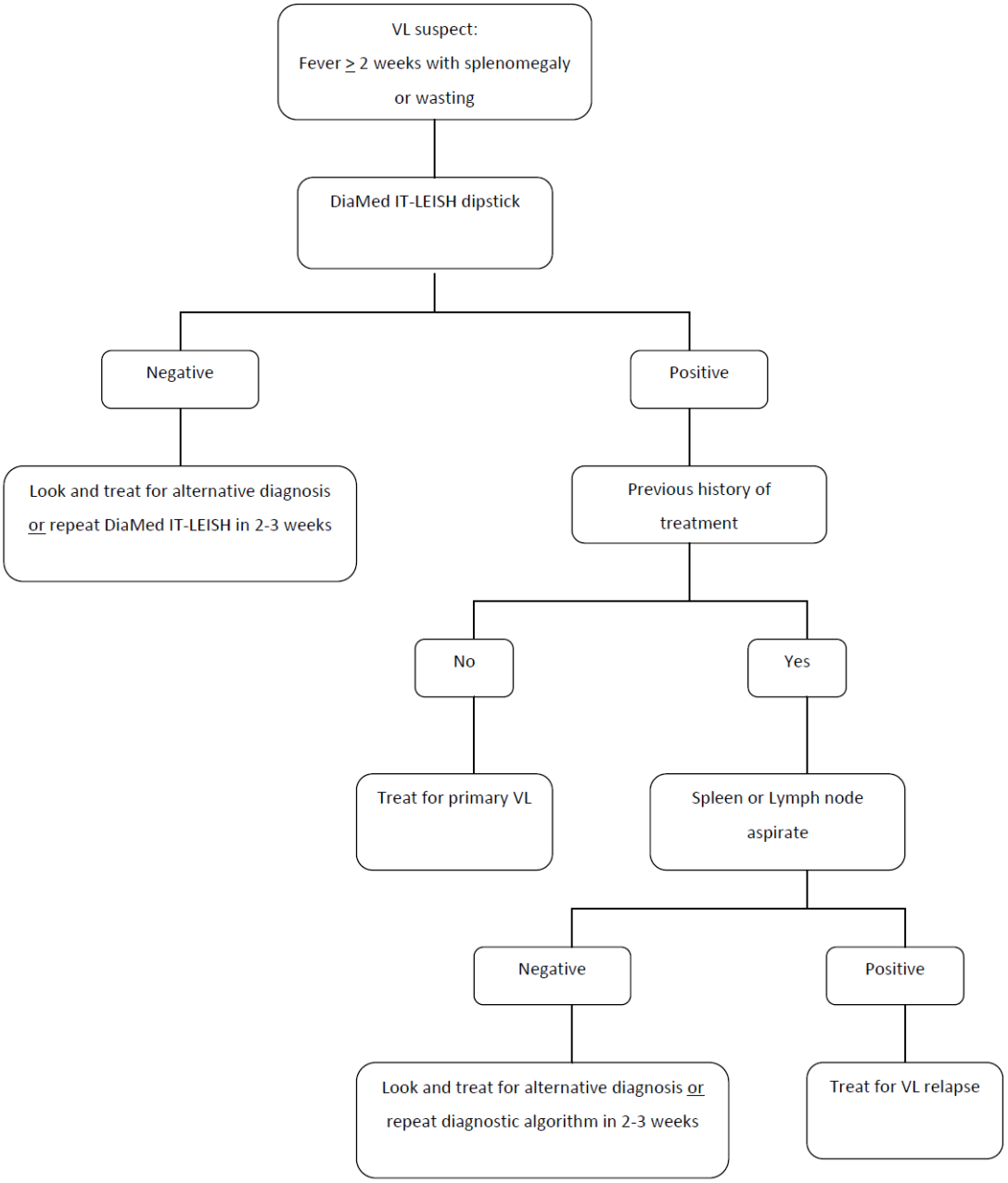

Figure 1 Diagnostic Procedure for VL [8]

5) First-Line Treatment Regimens for Primary VL:

5.1) Combination Therapy: Sodium Stibogluconate (SSG) & Paromomycin: In the combination therapy, sodium stibogluconate (20mg/kg body weight/day), and paromomycin (15mg/kg body weight/day) injections are given intramuscularly for 17 days [11].

5.2) *Sodium Stibogluconate or Meglumin Antimoniate (Monotherapy):* Pentavalent antimonials (SbV) are available in two chemical forms: sodium stibogluconate (SSG) and meglumine antimoniate (MA) [2, 6, 8].

SSG in monotherapy is administered as intramuscular injection of 20mg/kg/day for 30 days either intramuscular or by a slow intravenous infusion within 5 minutes. The dose depends on the weight of the patient but the minimum recommended dose of SSG is 2ml. It is also recommended to split the dose of it is more than 10ml [8, 11].

In patients with severe ascites and/or oedema, the dose of SSG should be decreased by subtracting 5 kg (if weight > 40 kg), 2 kg (if weight 25 - 40 kg) or 1 kg (if weight 10 - 25 kg) from the patient's body weight. The minimum dose is 2 ml (200 mg) for children weighing less than 10 kg [8].

5.3) *Liposomal Amphotericin B (LAmB, AmBisome):* In East Africa, it is less effective and therefore requires higher doses; it is used in patients for whom SSG/PM is contra-indicated or ineffective [6].

Liposomal Amphotericin B is recommended in those patients with pregnancy, HIV-co-infection, severe illness, severe anemia, severe malnutrition and extremes of age (below 2 years or above 45 years) [11].

6) *Second-Line Treatment for Primary Visceral Leishmaniasis:* indications for the use of second-line VL treatment are drug toxicity, relapse, treatment failure, very severe illness, pregnancy and VL/HIV co-infection [11].

6.1) *Liposomal Amphotericin B (AmBisome):* this is an extremely safe and efficacious drug for the treatment of visceral Leishmaniasis. The recommended dose of AmBisome for treatment is 5mg/kg/day over a period of 6 days (i.e. 30mg/kg in total) [11].

6.2) *Miltefosine:* this is the only oral anti-leishmanial drug taken at a dose of 2-3mg/kg per day (100mg/day for patients weighting more than 25kg) for 28 days. In one study in Ethiopia, Miltefosine was found to be less effective but safer than antimonials in HIV/VL co-infected patients [6, 8, 11].

6.3) Paromomycin (Aminosidine): a PM is used as a highly effective first line treatment in combination with SSG. The 15mg/kg sulphate is equivalent to 11mg/kg base, whereas 20mg/kg sulphate is equivalent to 16mg/kg base. Pain at the injection site is the most common adverse event [2, 6, 11].

7) Treatment of VL Relapse: a patient who is diagnosed with visceral Leishmaniasis for the first time is called a primary VL case and If a person returns with clinical features and a positive parasitology consistent with visceral Leishmaniasis, after having been successfully treated for primary VL and discharged improved or with a negative test of cure (TOC), the patient is known as relapse VL case [11].

The following regimens are recommended for the treatment of visceral Leishmaniasis relapses [11]:

7.1) *Liposomal amphotericin B:* 5mg/kg/day for a period of 6 days. The treatment period can be extended up to 14 doses for better therapeutic advantages.

7.2) *SSG*: 20 mg/kg/day for 40days until 2 consecutive weekly aspirates for parasitology is negative. If TOC is still positive, SSG should be given for a total of 60 days with close monitoring of drug toxicity and TOC checked again, if positivity persists, then 2nd line treatment must be used.

7.3) *SSG/PM combination:* SSG (30 days) and Paromomycin (17 days) can be used for VL relapse treatment.

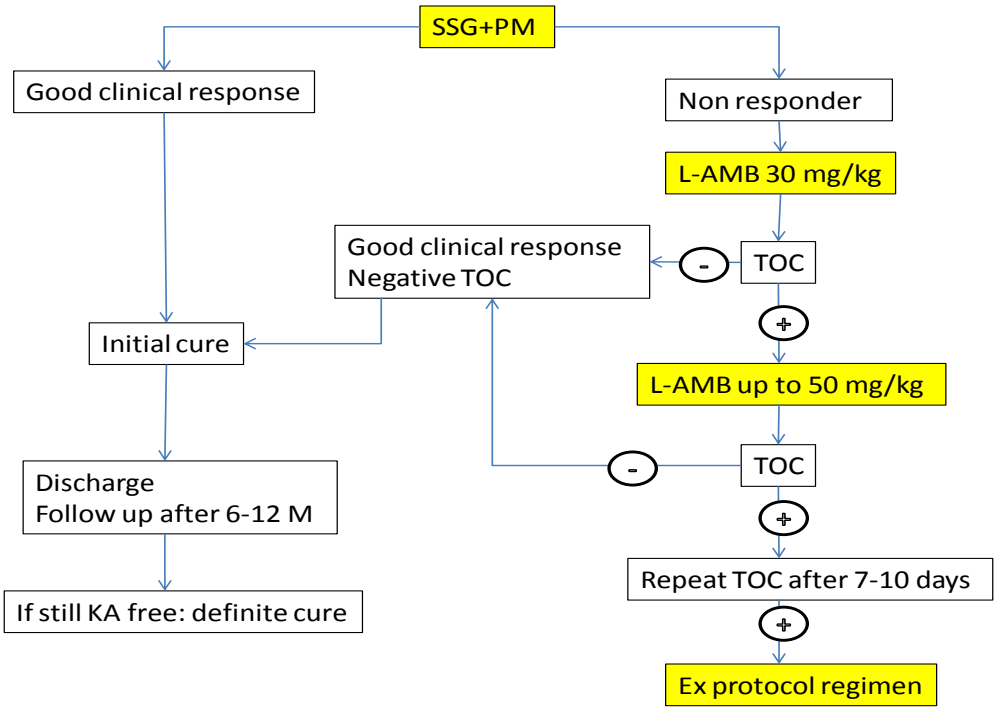

Figure 2 Treatment Procedure for Primary KA in Africa [6]

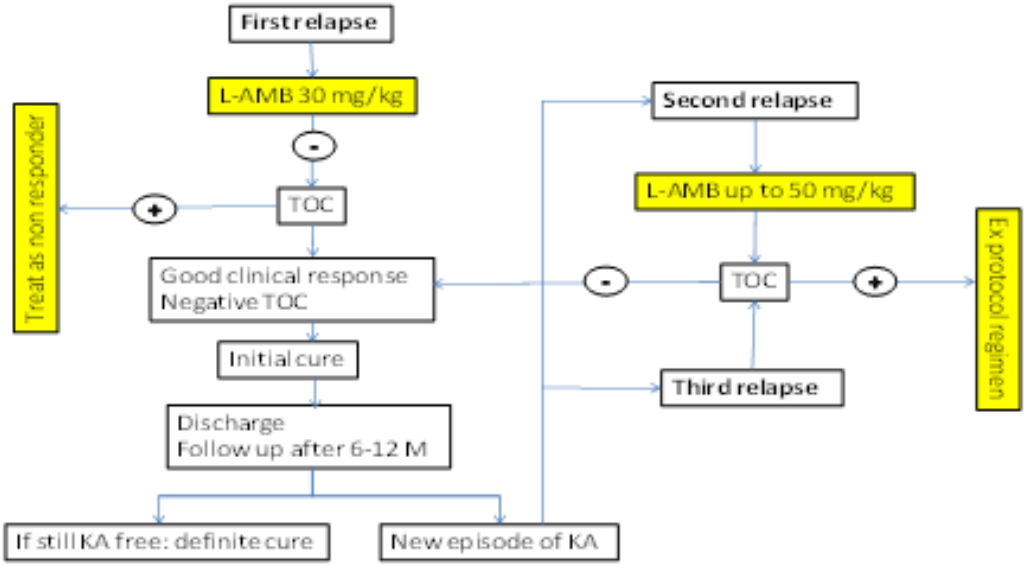

Figure 3 Treatment Procedure for Primary KA in Africa [6]

B. Automatic Knowledge Acquisition (Data Mining)

Nowadays, data stored in medical databases are growing in an increasingly rapid way. Due to this, tendency for data mining application in healthcare today is great, because healthcare organizations today are capable of generating and collecting a large amounts of data. This increase in volume of data requires automatic way for these data to be extracted when needed. With the use of data mining techniques it is possible to extract interesting and useful knowledge and these knowledge can be used by physicians to determine diagnoses, prognoses and treatments for patients in healthcare organizations which improve work efficiency and enhance quality of decision making process. In addition to this, Data mining tools can be very useful to control limitations of people such as subjectivity or error due to fatigue, and to provide indications for the decision-making processes [14, 17].

The data for this study have been collected from MSF Holland, Abdurafi project which is found in the North West Ethiopia and the organization have collected these data from 2005, 2006, 2008, 2009, 2010, 2011, 2012 and 2013 years which means that the organization does not have a data for the year 2007. During these years the organization has used different types of data base systems and the records that exist in these data bases contain different attributes in number and type. The original size of the whole data before pre-processing was 18.3 MB. The following table summarizes the number of attributes and number of records that are collected.

Table 1 Number of Attributes and Records

Year	Nº of Attributes	Nº of Records
2005	51	473
2006	51	182
2008	56	160
2009	80	246
2010	30	398
2011	76	336
2012	146	283
2013	146	604
Total		2682

1) *Data Cleaning:* Anticipating that data will be 100% complete and error free is unrealistic when working with patient data which collected in complex health care systems because Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data [13, 14].

We cleaned the data that has been collected from MSF Holland, Abdurafi project separately according to the year. We used maximum values and average values to handle missing attribute values. Since the 2010 data are significantly incomplete, we excluded the entire data from the data set. Since the data for the year 2005 and 2006 have been recorded for each month separately, first we integrated these data into one to make the data cleaning process more convenient and easy. Therefore, the data that have been recorded for each month in 12 sheets have integrated into one sheet.

The organization has used different codes for "Treatment site", "Residency status", "Exit code", "Sex", "Prior episode of VL", "Admission: Weakness", "Admission: Jaundice", "Admission: Oedema/Ascites", "DAT outcome", "Final diagnosis", "Final treatment", for the year 2012 and 2013 data and the researcher replace them accordingly.

2) *Data Integration:* Data integration combines data from multiple sources into a coherent data store. There are a number of issues to consider during data integration. Schema integration and object matching can be tricky. How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the entity identification problem. For example, how can the data analyst or the computer be sure that customer-id in one database and cust-number in another refers to the same attribute?[13]

Redundancy is another important issue. An attribute (such as annual revenue, for instance) may be redundant if it can be "derived" from another attribute or set of attributes. A third important issue in data integration is the detection and resolution of data value conflicts. For example, for the same real-world entity, attribute values from different sources may differ. This may be due to differences in representation, scaling, or encoding. For instance, a weight attribute may be stored in metric units in one system and British imperial units in another [13].

Therefore based on the above principles, we conducted data integration as follows: Since the data set has different attributes for different years, we integrated the data by taking the common attributes which are found in all years' data and we selected 27 common attributes for all year. Since the organization has used different database types for different years, there are two attributes which the organization used to store data about how long the patient stay in that area. These attributes are "Months in area" and "Years in area". Due to this, we converted months in the area to years in the area to avoid data value conflicts. The organization has used two different attribute names ("Discharge Date" and "Treatment End date ") to register the date on which the patient left the hospital and to avoid entity identification problem, we used "Discharge Date" while integrating the data.

There are two format types for the attribute values of "Months in area" attribute which are in month and in year and we changed these formats into one format which is month. The organization has also used two type of date format (MM/DD/YY and DD-MM-YY) to enter the attribute values for "Admission date" and "Treatment End date" attributes in the year 2005 and 2006 records and we changed these two format into one format which is MM/DD/YY.

The organization has used two attribute values ("Migrant" and "Migrant Worker") which have similar meaning to enter an attribute value for "Type of Residency" attribute because of this the we used "Migrant" and replace the reset of the attribute values with it to avoid inconsistency. The organization used two attribute values ("Primary" and "Primary KA") to enter an attribute value for "Diagnosis" attributes for the year 2005 and 2006 but both of them have the same meaning and due to this we changed "Primary KA" to "Primary" to avoid inconsistency.

Since the attribute values for "Treatment End date" and "Discharge Date" attributes in the year 2008 are similar, we removed the "Discharge Date" attribute from the record to avoid redundant data. There are two attributes (Age (years) and Age) in the year 2012 and 2013 data which have the same data and due to that we removed the former one to avoid redundancy.

3) *Data Transformation:* Data Transformation techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data.This leads to aconcise,easy-to-use,knowledge-levelrepresentationofminingresults [13].

Since the attribute values of ''Age''and ''BMI" are continuous and various, we used data transformation and replace the actual data with the result to make the data more suitable for data mining according to the following ranges:

Table 2 Data Transformation

Attribute	Range	Distnict Values before data discretization	Distnict Values after data discretization
Age	0-2=Very young, 2-45=Adult and >45=Old	67	3
BMI	0-18=Under Weight, 18-25=Normal and >25=Over Weight	547	3

4) *Class Imbalance:* The class imbalance problem is prevalent in many applications, including: fraud/intrusion detection, risk management, text classification, and medical diagnosis/monitoring, etc. It typically occurs when, in a classification problem, there are many more instances of some classes than others. In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. Particularly, they tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class [18].

A number of solutions to the class-imbalance problem were proposed both at the data and algorithmic levels. At the data level, these solutions include many different forms of re-sampling such as over-sampling and under-sampling. SMOTE (Synthetic Minority Over-sampling Technique) is an over-sampling approach which generates synthetic examples in a less application specific manner. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors [18].

As we can observe in the figure below, the classes are imbalance and to avoid this, we SMOTE the data set and Discretize it before conducting the experiments and as a result the dataset increases from 2284 records to 7594 records.

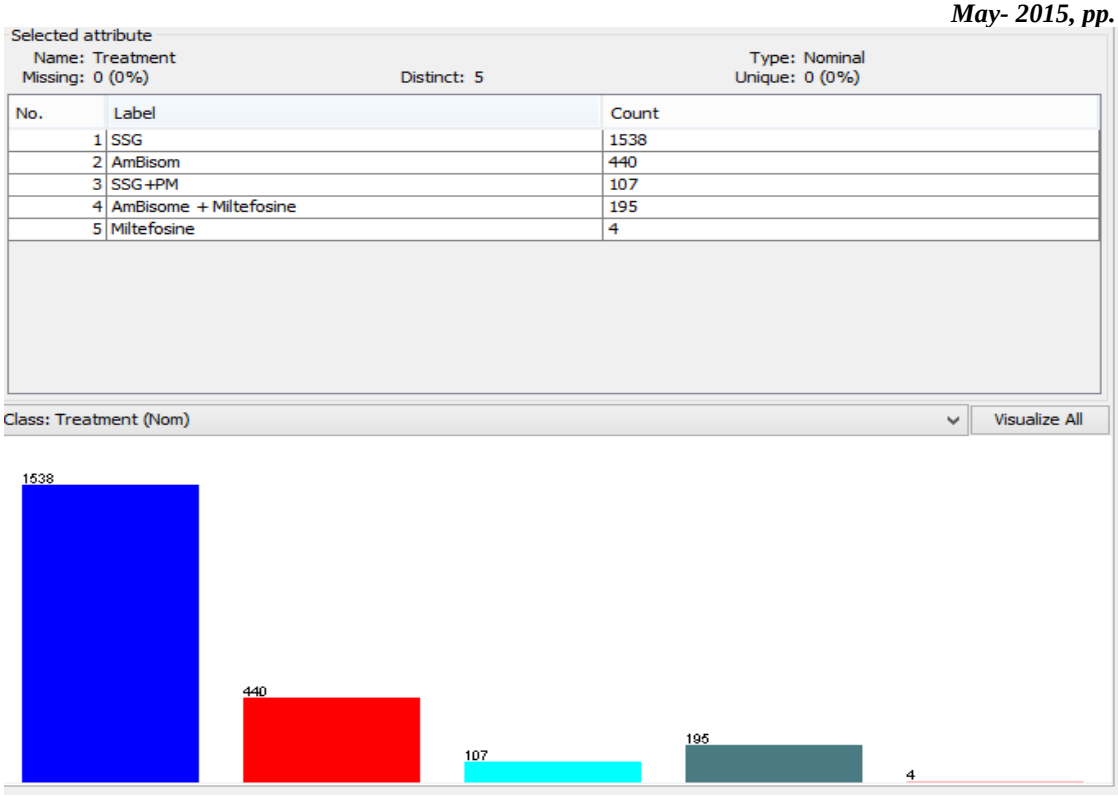

Figure 4 Imbalance classes before SMOTE

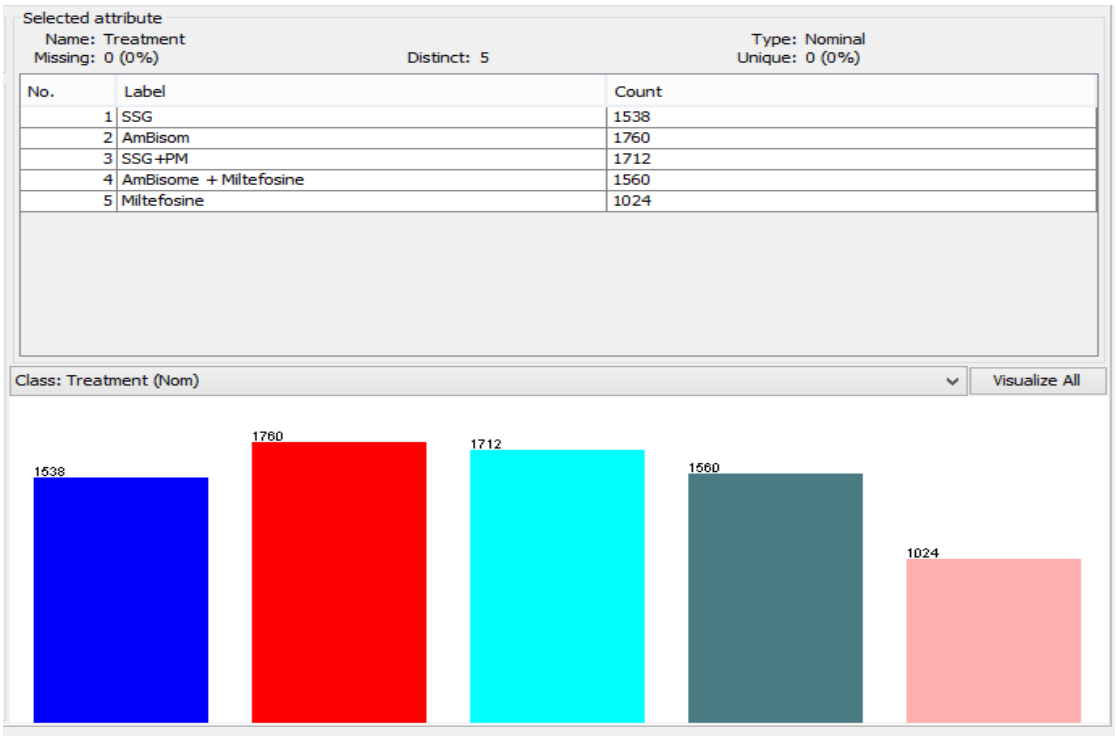

Figure 5 Balanced classes after SMOTE and Discrtize

5) *Attribute Selection:* Medical data is often very high dimensional. Depending upon the use, some data dimensions might be more relevant than others. In processing medical data, choosing the optimal subset of features is such important, not only to reduce the processing cost but also to improve the usefulness of the model built from the selected data [14]. The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit of reduceing the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand [13].

To select the best attibutes for data mining, we used information gain method which exist in WEKA data mining tool and the domain experts‘ advice and come up with 7 attributes which are listed here below:

Table 3 Attribute Description

S.No	Attibute Name	Value Type	Description
1.	Age	Nominal	Describes the patient's age
2.	Admission weakness	Nominal	Describes to what extent the patient is weak while s/he comes to hospital
3.	Body Mass Index	Nominal	Describes the patient's Body Mass Index measurement
4.	Admission Oedema	Nominal	Describes whether the patient has oedema or not while s/he comes to hospital
5.	Sero Statu	Nominal	Describes whether the patient is infected by HIV
6.	Diagnosis	Nominal	Describes whether the patient is infected by VL for the first time or more than once
7.	Treatment	Nominal	Describes what type of treatment the patient takes

6) *Data Mining:* Data mining refers to the application of algorithms for extracting patterns from data. Data mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/ evaluation step of the KDD process [14].

The objective of this step is to apply three classification technique algorithms on medical data set which have been collected from MSF Holland, Abdurafi Proejct and develop a model that can predict which treatment drug the patient should take from the available five options.

Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to the classes [13, 14].

We conducted three experiments for three classification algorithms namely J48 pruned, PART, and JRip under ten-fold Cross-Validation test option/mode. The data set has five classes namely SSG, AmBisom, SSG+PM, AmBisome + Miltefosine, and Miltefosine.

6.1) *Experiment one using J48 Pruned:* Decision tree is a graphical representation of the relations that exist between the data in the database. It is used for data classification. The result is displayed as a tree, hence the name of this technique. Decision trees are mainly used in the classification and prediction. It is a simple and a powerful way of representing knowledge [17].

This experiment conducted with default parameters of WEKA and the algorithm generates a model as a decision tree with 35 Number of Leaves and 52 Size of the tree and Correctly Classified Instances are 5131 which means 67.5665 % and Incorrectly Classified Instances are 2463 which means 32.4335 % from Total Number of Instances of 7594. The algorithm takes 0.02 seconds to develop the model.

Table 4Confusion Matrix for J48 classification algorithm

Confusion Matrix					
a	B	c	d	e	classified as
606	134	757	30	11	a = SSG
70	650	588	370	82	b = AmBisom
3	20	1689	0	0	c = SSG+PM
0	9	69	1162	320	d = AmBisome + Miltefosine
0	0	0	0	1024	e = Miltefosine

6.2) *Experiment two using PART:* PART is a rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form IF condition THEN conclusion.The "IF"-part (or left-hand side) of a rule is known as the rule antecedent or precondition.The "THEN"-part (or right-hand side) is the rule consequent [13].

This experiment conducted with default parameters of WEKA and the algorithm generates a model with 21 rules and Correctly Classified Instances are 5128 which means 67.527 % and Incorrectly Classified Instances are 2466which means 32.473 % from Total Number of Instances of 7594. The algorithm takes 0.07 seconds to develop the model.

Table 5 Confusion Matrix for PART classification algorithm

Confusion Matrix					
a	B	c	d	e	classified as
603	136	759	30	10	a = SSG
68	651	588	370	83	b = AmBisom
3	20	1689	0	0	c = SSG+PM
1	9	69	1161	320	d = AmBisome + Miltefosine
0	0	0	0	1024	e = Miltefosine

6.3) *Experiment three using JRip:* JRipis also a rule-based classifier uses a set of IF-THEN rules for classification and this experiment conducted with default parameters of WEKA and the algorithm generates a model with 8 rules and Correctly Classified Instances are 5109 which means 67.2768 % and Incorrectly Classified Instances are 248 which means 32.7232 %from Total Number of Instances of 7594. The algorithm takes 0.28 seconds to develop the model.

Confusion Matrix					
a	b	c	d	e	classified as
584	152	759	33	10	a = SSG
69	648	588	373	82	b = AmBisom
1	22	1689	0	0	c = SSG+PM
0	6	69	1164	321	d = AmBisome + Miltefosine
0	0	0	0	1024	e = Miltefosine

7) *Evaluation:* We used both objective and subjective interestingness evaluation methods.Objective interestingness measurement is generally based upon the inherent structure of mined patterns, i.e., the patterns' statistics like support or confidence. Patterns might be considered interesting when they represent strong regularities, rare exceptions, or when they help to distinguish different groups of items etc [19].

Table 6 Objective evaluation results

	Classification algorithms		
Objective evaluations	J48 pruned	JRip	PART
Correctly Classified Instances	67.5665 %	67.2768 %	67.527 %
Incorrectly Classified Instances	32.4335 %	32.7232 %	32.473 %
Time taken per second	0.02	0.28	0.07
TP Rate	0.676	0.673	0.675
FP Rate	0.085	0.086	0.085
Precision	0.738	0.733	0.737
Recall	0.676	0.673	0.675
F-Measure	0.651	0.647	0.651

As we can observe from the table above, J48 classification algorithm performs better in all objective interestingness evaluation methods than the others with a slight difference. However, we can understand from the above table that there are 32% inccorect rules that are generated by the three algorithmes. Therefore, in order to identify and exclude these incorrect rules from the knowledge base, it is also better to use subjective interestingness evalaution method, which rely on some formalization of expectations or previous knowledge, because objective measures do not make use of the human analyst's background knowledge about the application domain [9]. To do so, we converted the rules that are extracted by three algorithms in a form which is easily understandable by human and gives them to the domain experts to evaluate the correctness of rules in accordance with their knowledge what they have before.

we distributed rules of the three algorithms for ten domain experts and collects eight of them and the table here below shows subjective interestingness evaluation results:

Table 7 Subjective interestingness evaluation results

	Classification algorithms		
Subjective evalaution	J48 pruned	Jip	PART
Domain Expert Knowledge	65.75%	62.5%	75.37%
No of rules	35	8	21

Based on both objective and subjective interestingness evaluation methods result, we decided to use rules that are generated by PART classification algorithm model for further use in the development of the knowledge base of the knowldge base system because it registered better performance than others and the rules are presented here below:

Table 8 Rules extracted by PART Classification algorithm

S.No	Rules
1.	Diagnosis = VL Relapse AND Sero status = Negative AND Admission Weakness = OK AND BMI = UnderWeight: AmBisome + Miltefosine
2.	Diagnosis = VL Relapse AND Age = Adult AND Admission Weakness = Severe AND Sero status = Positive AND Admission Oedema = No: AmBisome + Miltefosine
3.	Diagnosis = VL Relapse AND Sero status = Positive AND Age = Adult AND BMI = UnderWeight AND Admission Oedema = No AND Admission Weakness = OK: Miltefosine
4.	Sero status = Not known AND Admission Weakness = OK AND Age = Adult: SSG
5.	Sero status = Discordant AND BMI = UnderWeight: AmBisom
6.	Sero status = Not known AND BMI = UnderWeight AND Age = Adult: SSG
7.	Sero status = Not known AND Age = Adult: SSG
8.	Sero status = Positive AND Diagnosis = VL Relapse AND Admission Oedema = No: AmBisome + Miltefosine
9.	Sero status = Not known AND Age = Very Young: AmBisom
10.	Sero status = Positive AND Diagnosis = Primary KA AND Age = Adult: AmBisome + Miltefosine
11.	Admission Weakness = Severe AND Sero status = Negative AND Admission Oedema = No AND Age = Adult: AmBisom
12.	Admission Weakness = Severe AND Sero status = Negative AND Age = Adult: AmBisom
13.	Sero status = Positive AND Diagnosis = Primary KA: AmBisom
14.	Age = Very Young: AmBisom
15.	Sero status = Not known: SSG
16.	Admission Oedema = Yes AND Sero status = Negative AND BMI = UnderWeight: AmBisom
17.	Sero status = Positive AND Admission Oedema = Yes: AmBisom
18.	Sero status = Negative AND Age = Old AND Admission Weakness = OK: AmBisom
19.	Sero status = Negative AND BMI = Normal AND Admission Oedema = No: SSG+PM
20.	Sero status = Negative AND Diagnosis = Primary KA AND Admission Weakness = OK AND Admission Oedema = No: SSG+PM
21.	: SSG

IV. KNOWLEDGE REPRESENTATION

Rule-based knowledge representation method was used because it is predominant means of representing the vast amount of problem specific knowledge in KBS in the form of 'situation » action', i.e. 'IF a certain situation holds THEN take a particular action' [4, 9]. Besides to this, the nature of the disease forced the researchers to use rule-based knowedge representation method because the knowledge that are acquired from data mining were rules and the guidelines of diagnosis and treatment of VL are full of rules and decision trees which could be easily converted to rules.

Rule 1 for Clinical Diagnosis

IF the patient has

Chronic fever for more than two weeks AND Weight loss

AND Spienomegaly AND Hepatomegaly AND

Wasting AND Darkening skin AND Dry cough AND
Bleeding AND Appetite loss AND Tiredness AND Travel history
THEN
VL Suspected=Yes.

Rule 2 for Clinical Diagnosis

IF the patient has
Chronic fever for more than two weeks AND Weight loss AND Spienomegaly AND Travel history
THEN
VL Suspected=Yes.

Rule 3 for Parasitologic Laboratory Diagnosis

IF the patient's
White blood cell decreases AND hematocrite decreases AND Platelet count decreases AND Parasitologic Diagnosis is positive
THEN
The patient is VL infected.

Rule 4 for Serologic Laboratory Diagnosis

IF the patient's
White blood cell decreases AND hematocrite decreases AND Platelet count decreases AND Serologic Diagnosis is positive
THEN
The patient is VL infected.

Rule 5 for Antibody Detection Laboratory Diagnosis

IF the patient's
White blood cell decreases AND hematocrite decreases AND Platelet count decreases AND Antibody Detection is positive
THEN
The patient is VL infected.

Rule 6 for Antigen Detection Laboratory Diagnosis

IF the patient's
White blood cell decreases AND hematocrite decreases AND Platelet count decreases AND Antigen Detection is positive
THEN
The patient is VL infected

Rule 7 for Treatment Ordering

IF the patient's
Diagnosis = VL Relapse AND Sero status = Negative AND Admission Weakness = OK AND BMI = UnderWeight
THEN
The patient should take AmBisome + Miltefosine.

Rule 8 for Treatment Ordering

IF the patient's
Diagnosis = VL Relapse AND Age = Adult AND Admission Weakness = Severe AND Sero status = Positive AND Admission Oedema = No
THEN
AmBisome + Miltefosine.

Rule 9 for Treatment Ordering

IF the patient's
Sero status = Negative AND Diagnosis = Primary KA AND Admission Weakness = OK AND Admission Oedema = No
THEN
SSG+PM.

V. IMPLEMENTATION

All the knowledge acquired through integrated knowledge acquisition techniques represented accordingly in the knowledge base using SWI-Prolog 6.4.0 with UTF-8 which allowed us to develop the Knowledge Based System in Amharic Language. Besides to this, Java NetBeans IDE 7.3 with JDK 6 was employed to integrate WEKA result with the Knowledge Based System automatically without domain experts interference. In addition to this, we have developed the GUI of the proposed system using Java.

As we can see from Figure 6 here below, the proposed system has five components such as Knowledge Base, Inference Enginee, User Interface, Explanation Facility, and Knowledge Base Editor. Knowledge Base stores all the knowledge, in the form of rules, acquired from manual and automated knowledge acquisition techniques which are required for diagnosis and treatment of VL.

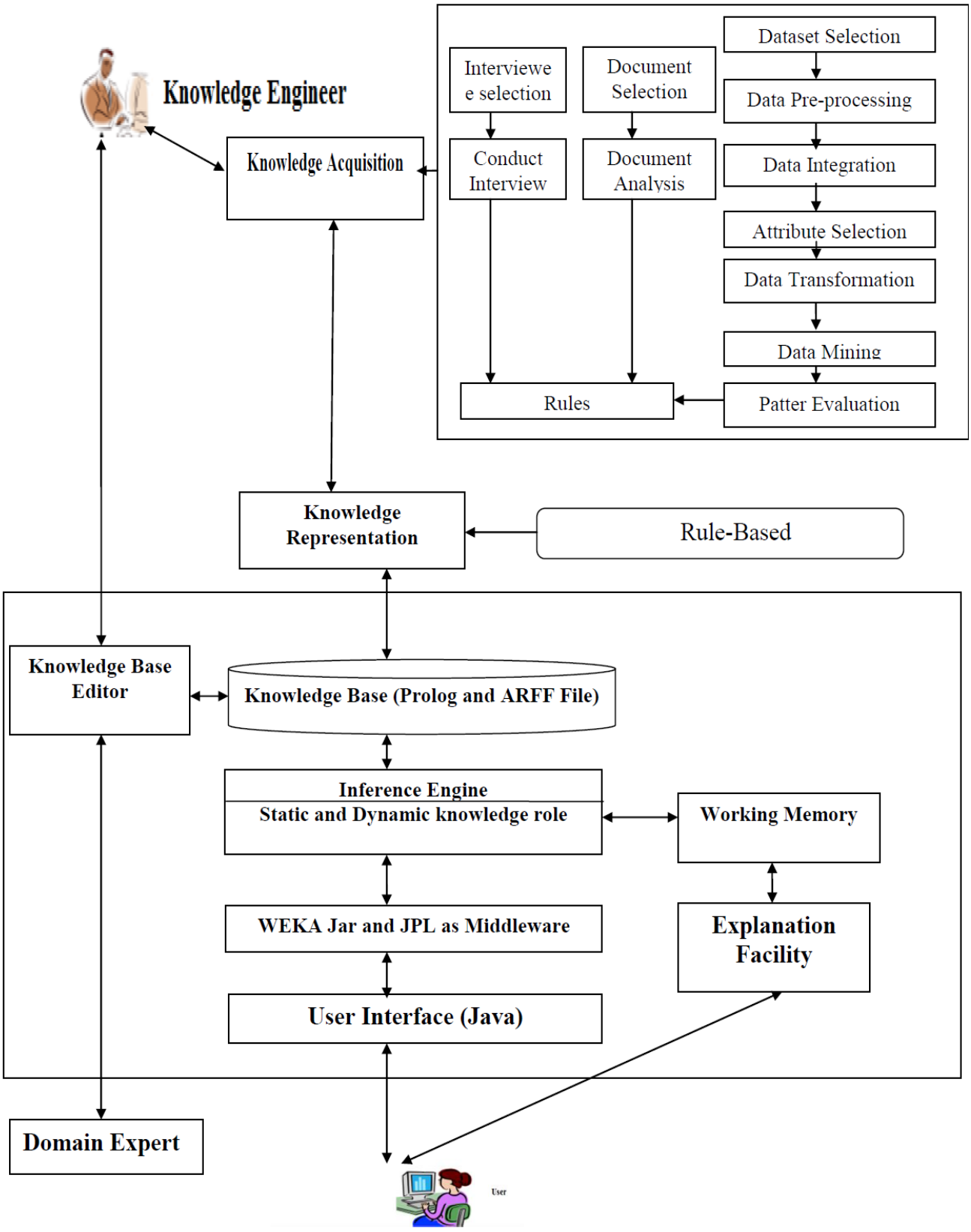

Figure 6 Architecture of the proposed KBS for Treatment and Diagnosis of VL

An inference engine is the brain of the Knowledge Based System which directs the system how it can derive a conclusion by looking for possible solutions from the knowledge base. Since the objective of the proposed KBS is to diagnosis and treatment of VL and the Prolog's built-in inference mechanism is backward chaining, the researchers prefered to use backward inference mechanism which is a goal derive that tries to prove or disprove the goal.

The user interface facilitates the communication between the system and the user. When the user wants to use the proposed system, s/he has to run the prototype. Then after the following options 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.

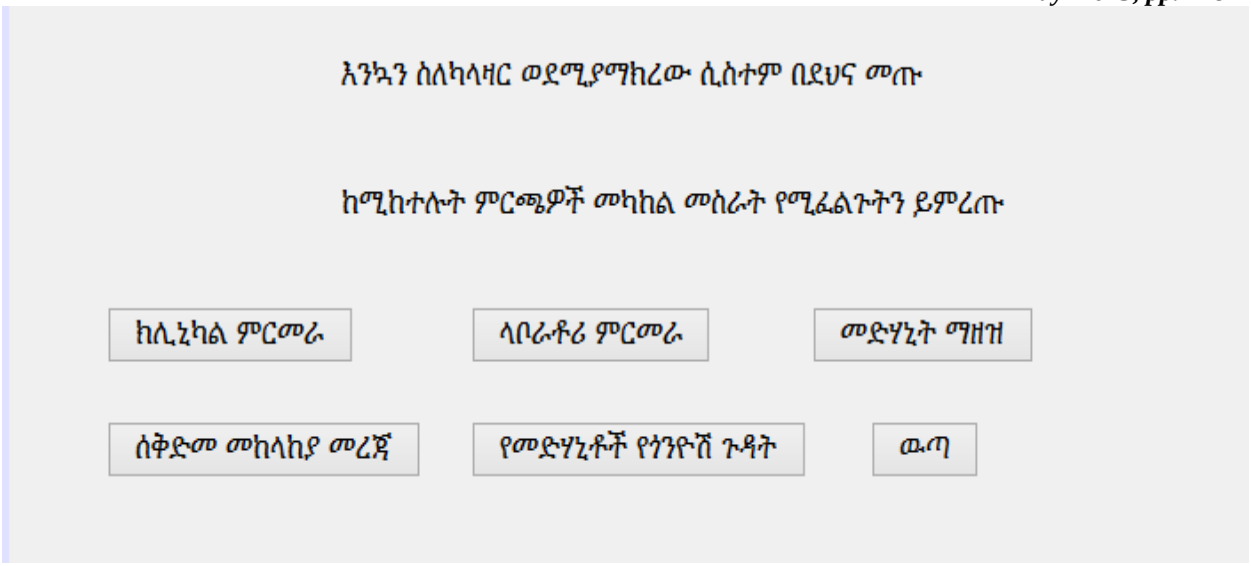

Figure 7 Home Graphical User Interface of the proposed system

As we can observe from the Figure 7 here above, the user has five options. To enter her/his option from the available five options, the only thing that the user has to do is double click on the button and then after based on the user choice the system displays the next step. For instance, if the user wants to conduct clinical diagnosis, Figure 8 will be displayed as follows:

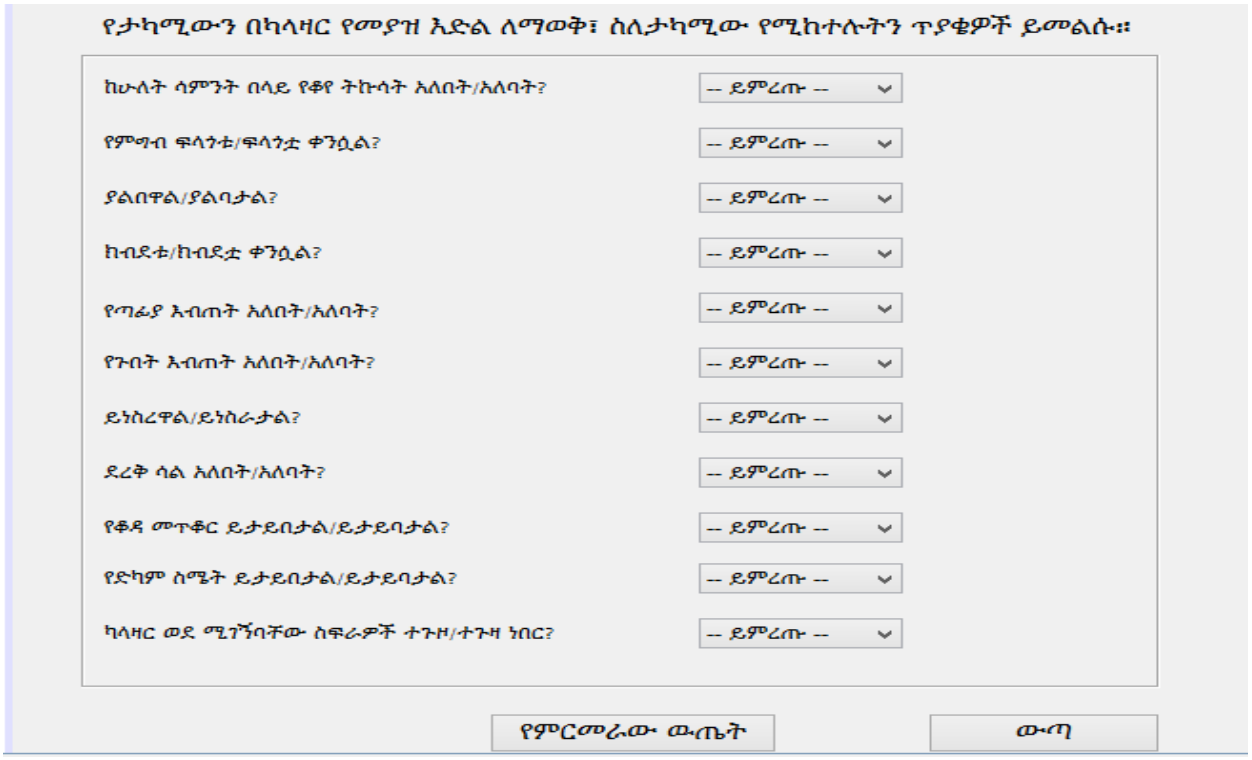

Figure 8 Graphical User Interface for Clinical Diagnosis

The other component of the proposed system is the explanation facility which describes how the system reaches on conclusion based on the set of facts and rules that are available in the knowledge base. The proposed system supports its conclusion with explanation. In addition to this, the explanation facility component of the proposed system gives additional information for the users about the dosage of VL treatments and their side effects as well.

Knowledge Base Editor is one of peripheral features of the proposed KBS through which the knowledge engineer in collaboration with the domain expert can edit the knowledge base whenever there is an update, because knowledge is dynamic, which has to be included in the knowledge base.

Finally, in addition to addressing the problem mentioned, the contribution of this study in the general architecture of the Knowledge Based System is the knowledge acquisition techniques which means that in this study integrated (manual and automated) knowledge acquisition techniques have been implemented. Besides to this, the data mining results are integrated automatically using WEKA Jar file and JPL jar has also used use the knowledge base that has been constructed by prolog.

VI. EVALUATION AND RESULTS

A. System Performance Testing Using Test Cases

Test cases are one of the predominant evaluation mechanisms for evaluating the performance of the proposed system which helps to compare and contrast the domain experts' judgment and the proposed system's response so that conclusions could be made on whether the proposed system could work in the absence of domain expert or not. To do so, 20 test cases were prepared in collaboration with the domain experts and given to the proposed system and domain expert. As a result the proposed system could perform in the absence of domain experts with 95% which indicates that the study was effective in acquiring the required knowledge for diagnosis and treatment of VL through integrated (manual and automated) knowledge acquisition techniques.

B. User Acceptance Testing

The other way of evaluating the proposed system is user acceptance testing through which whether the potential users would like to use the proposed system frequently or not could be evaluated. Then user acceptance testing questionnaire were prepared which the researchers adopted from Measuring Usability with the System Usability Scale (SUS), which was released into this world by John Brooke in 1986 and has become an industry standard with references in over 600 publications[10], with slight modification. The questionnaire has ten close-ended questions. As a result, the proposed system achieved 86% of the user acceptance which is the promising result to implement the proposed system in remote areas where the VL is endemic and conduct further researches on this area. Besides to this, 87% and 13% of the evaluators Strongly Agree and Agree, respectively, that the proposed system could help in the effort to improve the health care service in remote areas where senior health experts are not available.

VII. DISCUSSION AND CONCLUSION

As we have discussed in the evaluation section, the proposed system achieved promising results with 95% system performance testing and 86% user acceptance testing results which are the better results as compared to [1] which achieved 80.5% overall performance by using only data mining as means of knowledge acquisition technique and to [7] which achieved 84.2% overall performance by using only Interviews and documents analysis as means of knowledge acquisition technique which in turn indicates that using integrated (manual and automated) knowledge acquisition techniques is better than using manual or automated knowledge acquisition techniques separately in constructing the Knowledge Based System. However, while we did the data mining process, we noticed that there are notes that have been written by health experts as remark. So that, if text mining could be applied, new cases or knowledge could be discovered such that we could expand the knowledge base and improve the performance of the system.

ACKNOWLEDGMENTS

We would like to use this opportunity to express our gratitude to MSF Holland, Abdurafi Project for giving us the dataset without which the data mining part of this study would be impossible. In addition to this, we would like to gratitude everyone who works in Leishmaniasis Research and Treatment Center of University of Gondar Hospital for their willingness to participate in this research.

REFERENCE

[1] Johan van Griensven and Ermias Diro. "*Visceral Leishmaniasis*". Infectious Disease Clinics of North America, Vol 26, Issue 2, PP 309-322, Jun. 2012.

[2] Abdulkerim Mohammed. "*Towards Integrating Data Mining with Knowledge Based System: The Case of Network Intrusion detection*". M.Sc. Thesis, Addis Ababa University, June, 2013.

[3] Médecins Sans Frontières (MSF). "*Deadly if untreated: pushing forward the response to kala azar in Ethiopia*".

[4] RAMANI, R. V. and PRASAD. K. V.K. "*Applications of knowledge based systems in mining engineering*." APCOM 87. Proceedings of the Twentieth International Symposium on the Application of Computers and Mathematics in the Mineral Industries. Volume 1: Mining. Johannesburg, SAIMM. 1987. pp. 167- 180.

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

[6] MSF Holland *"Manual for the Diagnosis and Treatment of Visceral Leishmaniasis (Kala Azar) Under Field Conditions"*, MSF-Holland, 2014.

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

[8] Ministry Of Health Uganda." *The Diagnosis, Treatment and Prevention of Visceral Leishmaniasis in Uganda*". Guidelines for Clinicians and Health Workers, April 2nd 2007.

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

[10] Jeff Sauro." *Measuring Usability with the System Usability Scale (SUS)*". February 2, 2011. Available: http://www.measuringu.com/sus.php [25 November 2014].

[11] Ethiopian Federal Ministry of Health."*Guideline for Diagnosis, Treatment and Prevention of Leishmaniasis in Ethiopia*".2 Edition. June, 2013.

[12] Marie José Vlaanderen. "*Automated Knowledge Acquisition for Expert Systems an Overview*." PhD dissertation, Faculty of Philosophy, Erasmus University Rotterdam, Netherlands, 1990.

[13] Jiawei Han and Micheline Kamber."*Introduction,*" In Data Mining: Concepts and Techniques, Second Edition, University of Illinois at Urbana-Champaign.

[14] Ing-song Li, Hai-yan Yu and Xiao-guang Zhang." *Data Mining in Hospital Information System*". Zhejiang University, China.

[15] Priti Srinivas Sajja, and Rajendra Akerkar. "Advanced Knowledge Based Systems: Model, Applications & Research", In TMRF e-Book, Vol. 1, pp 1 – 11, 2010.

[16] Prof. Mihaela OPREA PhD." *On the Use of Data-Mining Techniques in Knowledge Based Systems*", Department of Informatics, University Petroleum-Gas of Ploieşti.

[17] Boris Milovic and Milan Milovic. "*Prediction and Decision Making in Health Care using Data Mining*". International Journal of Public Health Science (IJPHS), Vol. 1, No. 2, December 2012, pp. 69~78.Nov 8, 2012.

[18] Chen, Yetian. "*Learning classifiers from imbalanced, only positive and unlabeled data sets*." Department of Computer Science, Iowa State University. 2009.

[19] Pohle, Carsten. "*Integrating and Updating Domain Knowledge with Data Mining*." VLDB PhD Workshop. 2003.

