ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

UNCOVERING KNOWLEDGE THAT SUPPORTS MALARIA PREVENTION AND CONTROL INTERVENTION PROGRAM IN ETHIOPIA

GELETAW SAHLE

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

UNCOVERING KNOWLEDGE THAT SUPPORTS MALARIA PREVENTION AND CONTROL INTERVENTION PROGRAM IN ETHIOPIA

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics

By

GELETAW SAHLE

JUNE 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

UNCOVERING KNOWLEDGE THAT SUPPORTS MALARIA PREVENTION AND CONTROL INTERVENTION PROGRAM IN ETHIOPIA

By

GELETAW SAHLE

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
Dr. Million Meshesa	Advisor(s),	_____	_____
Dr. Wakgari Deressa	Advisor(s),	_____	_____
_____	Examiner,	_____	_____

DECLARATION

The thesis is my original work, has not been presented for a degree in any other university and that all sources of materials used for this thesis have been acknowledged.

_____

Geletaw Sahle

June 2011

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

1. Million Meshesa (PHD) _____ ____/_____/___
2. Wakgari Deressa (PHD) _____ ____/_____/___

ACKNOWLEDGEMENT

First and foremost I would like to thank God and the holy mother, who made all things possible and granted me success in my thesis work and entire journey.

Next, I would like to express special thanks to Dr. Million Meshesha and Dr. Wakgari Deressa for their help in the role of a supervisor starting from the first day of my research. They encouraged me to achieve to the best of my ability. Without them, I could not have reached to this stage. In many ways they provided me clear direction and valuable comments for the development of this thesis.

I would like to thank Dr. Fatumata Nato Traore (WHO representative) and Dr. Worku Bekele (WHO national malaria program officer) for giving permission to access national malaria data.

I would like to thank my other handful person, Henok Kebede (Data Management and GIS support unit), in the role of supervisory panel in the way of my research starts from the beginning by providing relevant materials, malaria data and their comments. Specially again commenting on the knowledge that will obtain from the research finding.

I would like to end by saying thank you to a handful people. Thanks, to WHO (specially the receptionists) for their supports in my administration procedures and their friendship. Thanks to the members of national metrological and national mapping agency staffs for their assistance, suggestions, and criticisms. Finally, thanks also to my friends, who are beside me during the up and down periods.

Special gratitude is extended to my friend Alula Kiros for his valuable support by facilitating access to get data from WHO. Also, special thanks to my friends Zelalem Hailu and Habtom G/Egizihabiher for their facilitating access to get print paper and access respectively.

Last but not least I would like to thank my family who always striving for my comfort, friends and colleagues for their support and encouragement throughout this process, without them this would not have been possible.

ACRONYMS

AAU	Addis Ababa University
AIDS	Acquired Immuno-Deficiency Syndrome
ART	Antiretroviral Therapy
CRISP-DM	Cross Industry Standard for Data Mining
DHS	Demographic Health Survey
DM	Data Mining
DMT	Data Mining Technology
EDHS	Ethiopian Demographic Health Survey
Epi Info	Epidemiological information
FMOH	Federal Ministry of Health of Ethiopia
GIS	Geographical Information System
GMP	Global Malaria Programme
HIV	Human Immuno-deficiency Virus
IDS	Integrated Disease Surveillance
IPT	Intermittent Preventive Treatment
ITNs	Insecticide Treated Nets
KDD	Knowledge Discovery from Data
MDG	Millennium Development Goal
MLP	Multilayer Perceptron
OLAP	On-Line Analytical Processing
PF	*Plasmodium falciparum*
PLT	Analysis of Platelet
PV	*Plasmodium vivax*
SPSS	Statistical Package for Social Sciences
WHO	World Health Organization
UNICEF	United Nations Children's Fund (formerly United Nations International Children's Emergency Fund)
USAID	United States Agency for International Development

TABLE OF CONTENT **PAGES**

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Malaria is one of the leading causes of death in Ethiopia. Though there are many efforts to control malaria, the complexity of the problems is still very severe. So there is a need to investigate in detail the synergic effect of risk factors with temperature, altitude, type of visit and malaria type and their causes of death. Hence in this research an attempt is made to determine the hierarchical importance of different risk factors and their patterns on malaria death occurrence.

In this study, knowledge discovery techniques are evaluated to support and uncover knowledge to scale up the malaria prevention and intervention program in Ethiopia. CRISP methodology with classification algorithms such as J48, JRip and MLP and pattern discovery analysis techniques like Apriori adopted to uncover knowledge for the mining of interesting rule from total datasets of 37, 609 records. An attempt is made to preprocess the data using business and data understanding with detail statistical summary in order to fill missing and detect noisy value. Essential target dataset attributes have been constructed by integrating WHO malaria databases, National Metrological data and National Mapping Data.

All classification techniques discover important attributes/factors that determine the malaria type of cases and occurrence of deaths. J48 Decision tree and MLP correctly classify 95.9% and 97.4%, respectively to predict occurrence of death. The findings of this rsearch indicate that rainfall is the significant factor that determines the prevalence of malaria. When the number of malaria cases increases there is a probability of death occurrences; the risk is relatively high with those less than 5 years age. In most zones, malaria transmission rate is high from May to January because of favorable climate conditions for malaria reproduction.

Apriori techniques (both general and class association mining) also strengthen the result of J48 and MLP. More interestingly, it discovers occurrence of death, mostly related with severe anemia cases rather than pregnancy cases. Such interesting rule needs further investigation to validate.

CHAPTER ONE

INTRODUCTION

Malaria is caused by a parasite called Plasmodium, which is transmitted via the bites of infected mosquitoes. In the human body, the parasites multiply in the liver, and then infect red blood cells. Symptoms of malaria include fever, headache, and vomiting, and usually appear between 10 and 15 days after the mosquito bite. If not treated, malaria can quickly become life-threatening by disrupting the blood supply to vital organs. In many parts of the world, the parasites have developed resistance to a number of malaria medicines. Key interventions to control malaria include: prompt and effective treatment with artemisinin-based combination therapies, use of insecticidal nets by people at risk and indoor residual spraying with insecticide to control the vector mosquitoes [**1**].

1.1. Malaria and It's Burden in Ethiopia

Identifying public health issues of the community is one of the most important steps in the planning of health care interventions. However, appropriate planning of health programs in turn depends to a large extent on access to timely and accurate information on demographic characteristics, on the occurrence of major health problems, and on associations with underlying factors [**1, 2**].

Perlino **[2]** describes public health as professionals from many fields with the common purpose of protecting the health of a population. For example:- Public Health helps in policy and practice like vaccination programs for school-age children and adults to prevent the spread of disease, regulation of prescription drugs for safety and effectiveness, safety standards and practices to protect worker health and safety , educational campaigns to reduce obesity among children, ensuring access to clean water and air measurement of the effect of air quality on emergency recovery worker, school nutrition programs to ensure kids have access to nutritious food and so on **[2].**

Malaria is one of the leading problems (health issues) in Africa especially in sub-Saharan countries. Every year it is the leading cause of outpatient consultations, admissions and death [**1**]. Hereunder, we try to review the report that shows the severity of malaria in Africa in general and in Ethiopia in particular.

USAID [**3**] report shows there are 500 million cases of malaria occur every year, directly and indirectly causing more than 1 million deaths. Ninety percent of these deaths occur in Africa, and most of the victims are young children. Malaria is responsible for at least 20 percent of all deaths among children under age 5 in Africa. Malaria also places a huge burden on already fragile health systems, representing 30 to 50 percent of outpatient visits and hospital admissions and it continues a major underlying barrier to economic development in Africa [**3**].

Similarly, in Ethiopia, malaria is one of the leading causes of health burden and results immense economic destruction. The country fights against malaria more than half a century ago. Due to the severity of malaria deaths and cases the FMOH focus attention on reducing overall burden by implementing different polices and strategies [**5, 6, 8**]. Additionally UNIECF, WHO and other governmental and non governmental organization focus their attention very well on malaria by setting different global and local policies, strategies and consultancies to the support the FMOH. Table 1.1 summarized the report and survey result of different responsible organization in Ethiopia.

Table 1.1 Summarize malaria report in Ethiopia.

Publisher	Description
UNICEF Ethiopia [**4**]	Indicates malaria burden contributes up to 20% of under-five deaths. Tragically, in epidemic years, mortality rates of nearly 100,000 children are not uncommon. For example: the malaria epidemic in 2003, were up to 16 million cases of malaria - 6 million more than an average year and out of an estimated 9 million malaria cases annually, only 4-5 million has been treated in a health facility. It is estimated that only 20 per cent of children under five years of age that contract malaria are treated in a facility. The remainder will often have no medical support. PF and PV are two common malarial parasites in the region. The former is considered the most severe of the two and almost all deaths occur by infection from this parasite. PF can rapidly become resistant to malarial treatment and poses a significant challenge to malarial medicine

Ethiopia National Malaria Indicator Survey 2007 Technical Summary [**5**]	Clearly figure out malaria is seasonal with unstable transmission that lends to the outbreak of epidemics. The transmission patterns and intensity vary greatly depending on the diversity in altitude, rainfall, and population movement. The technical survey indicates areas below 2,000 meters are considered to be malarious which covers approximately 68% (52 million) of the Ethiopian population and cover almost 75% of the country's landmass
Federal Ministry of Health [**6**]	Indicates, malaria accounts for up to 17% of outpatient consultations, 15% of admissions and 29% of in-patient deaths. About 75% of the country is malarious (defined as areas <2000 m, those areas are fertile and suitable for agriculture), with about 68% (i.e. 50 million) of the country's total population living in areas at risk of malaria. Approximately 9.5 million clinical cases of malaria were reported annually between 2001 and 2005 (range: 8.4 – 11.5 million), with an annual average of 487,984 laboratory confirmed cases over the same period (range: 392,419 – 591,442). According to the report, approximately 70,000 people die of malaria each year in Ethiopia.
Demographic Health and Survey [**7**]	Strengthen the reports on [**1,3,4,5, 6]** and the disease is the primary cause of health problems with unstable transmission pattern and characterized by focal and cyclic large scale epidemics, accounting for 17 percent of outpatient visits, 15 percent of hospital admissions, and 29 percent of in-patient deaths

1.1.1. Malaria Action Taken In Ethiopia

Ethiopia establishes different responsible bodies or organization at national and regional levels that plan and implement policies and strategies to eradicate malaria transmission and burden as well as to improve the health of the people. For example: The major responsible organization in the history and current status of malaria control are:- malaria control organization in Ethiopia established in 1959 for malaria eradication service, vertical malaria control program was introduced in 1971 and

malaria control was integrated into general health system and decentralized occurred with in FMOH in 1993 [**6, 9**].

The Government of Ethiopia put its policy towards malaria control to gave priority to communicable diseases, free diagnosis (especially at lower health facilities level), free anti-malarial drugs, free distribution of Insecticide Treated Nets (ITNs) to all and free indoor residual spraying of houses **[9]**.

The fight against malaria is governed by a five-year strategic plan for 2006–2010 based on malaria control interventions that include distribution of insecticide-treated nets (ITNs), indoor residual spraying (IRS), and prompt and effective treatment with artemisinin-based combination therapy (ACT). Altitude, morbidity data, and history of epidemics are the main criteria to select the area and the selected villages within the malarious areas (below 2,000m) and scheduled to reach 60% of target areas by 2010 [**5**].

By considering the above burden Ethiopia recently developed a six year (2010-2015) national strategic plan for malaria prevention, control and elimination. The main goal is by 2015 the country achieve malaria elimination with specific geographical areas with historically low malaria transmission area and near zero transmission in the remaining malarious areas of the country [**6**].

All in all to prevent and protect the disease the country currently used different strategies like distribution of insecticide-treated nets (ITNs), indoor residual spraying (IRS), and prompt and effective treatment with artemisinin-based combination therapy (ACT) [5, 9]. Even if there exist such type of malaria prevention and control program in the country malaria is still severe and headache for the country because it is one of the top ten and the leading causes of outpatient visits and admission in the year 2009/2010 [**4, 6, 7,13**].

1.1.3. Importance of Mining Malaria Data

Myatt [**10**] said an unprecedented amount of data collected and generated from different field causes information overload and the ability to make sense this data requires an understanding of exploratory data analysis and data mining.

Hand and Kamber **[11]** explain data mining discovers interesting knowledge, regularities or high level information can be extracted from databases and viewed or

browsed from different angles that are applied to decision making, process control, information management, query processing and so on. Data mining considered as one of the most important frontiers in database systems and one of the most promising new database application in the information industry **[11]**.

A database is a store of information but more important is the useful information which can be inferred from it. In order to do this a wide variety of data-mining methods or techniques should be used. There is no particular rule that would tell you when to choose a particular technique over another one. Sometimes those decisions are made relatively arbitrary based on the availability of data mining analysts who are most experienced in one technique over another. These techniques can be used for either discovering new information within large databases or for building predictive models **[12]**. Decision tree, rule induction and neural networks are some of the prediction and classification techniques used in data mining.

There are also researches done in Ethiopia to investigate the application of data mining in health sector. For instance, Teklu [**14**] has conducted a data mining research to see its application on antiretroviral data. Shegaw [**15**] also integrated application of data mining technology to predict child mortality pattern, further Abraham [**16**] explored determinant factors of HIV infection. However, as to the researcher knowledge there is no research conducted to apply data mining for malaria.

1.2. Statement of the Problem and Justification

To combat malaria EFMOH [**7**] focuses to reduce malaria related deaths and illness by implementing different polices and strategies. For example: the goal of malaria prevention and control in Ethiopia is to contribute to MDG 6 target 8 by reducing the overall burden of malaria by 50% by the year 2010 and to contribute to the reduction of child mortality (MDG 4) and improvement of maternal health (MDG 5) [**7**]. To support this, USAID [**3**] also promote effective treatment of malaria illness, protect pregnant women from malaria, respond the emergence and spread of drug-resistant malaria, develop new tools and approaches for malaria prevention and control and addressing the needs of populations in complex humanitarian emergencies to control the disease.

Apart from the above effort, the underlining problem which forces this research is the leading causes of malaria severity in Ethiopia and complexity of the problems in

stopping and controlling the causes or deaths. As a matter of fact, the disease accounts top leading causes of outpatient visits in 2009/10 [**6**]. Remarkable efforts are being made in preventing and controlling the crisis of the disease. Some of the efforts made in reducing the death and cause of the disease has distribution of insecticide-treated nets (ITNs), indoor residual spraying (IRS), and prompt and effective treatment with artemisinin-based combination therapy (ACT) [**5, 9**]. Altitude, morbidity data, and histories of epidemics are also the main criteria to select the area and the selected villages within the malarious areas [**5**].

There are so many factors that complicate the malaria prevention and controlling delivery at the facility/community levels and become hot research area of many scholars. Junior and Duarte [**31**] investigate supporting tools for diagnosis of asymptomatic malaria using ANN and BN, Roca-Feltrer et al [**33]** identifies the relationship between transmission intensity, seasonality and the age pattern of malaria, Yé et al [**34**] assess the effect of meteorological factors on clinical malaria risk among children, Teklehaimanot et al [**35**] attempts to understand the reason for variation is crucial to determining specific and important indicators for weather prediction of *PF,* Newman et al [**36**] assess malaria burden in area of stable and unstable transmission during pregnancy and Roca-Feltrer et al [**39**] indicates malaria seasonality to aid localized policymaking and targeting of interventions.

Another unresolved issue is assessing the factors/patterns that affect occurrence of deaths and cases which is crucial for decision making related to malaria prevention plan. To elaborate this, counting the number of malaria admissions and cases from each district of the region using statistical tools may not be a right answer always. Rather there is a need to apply knowledge discovery tools in order to extract the hidden knowledge from existing data that guide and direct each region/district to come up with an effective plan for malaria intervention. At present, malaria prevention and controlling services are expanding in Ethiopia. Also studying what factors might affect the success of the program from the data collected during service delivery through powerful analysis tools has paramount importance to maintain continuity and promote good health of malaria infected people by preventing opportunistic infections and hence delaying death. The data may be too much to easily analyzed by the classical statistical methods as well as simple queries of database management system. Statistics traditionally is concerned with analyzing primary (e.g.

experimental) data that has been collected to check specific research hypotheses (`hypothesis testing'). As such statistics is `primary data analysis' or top-down (confirmatory) analysis. Therefore, data mining is a solution for discovering hidden but important patterns from large volume of data collected over time. In particular, data mining is known to be effective in dealing with the discovery of hidden knowledge, unexpected patterns and new rules from databases [**14**].

The development of malaria early warning systems to predict where and when malaria epidemics will occur, in order to target scarce resources for prevention activities has motivated many studies and results different determinants on malaria transmission reviewed well in [**37, 35**]. Also noted little consensus has emerged about the relative importance and predictive value of different factors. To fill this gap Teklehaimanot et al [**35**] model daily average number of cases using robust Poisson regression with climatic variables (rainfall, minimum temperature and maximum temperatures) that are predictors of transmission potential. However, the study doesn't consider the presence of some confounding factors associated with weather. Protopopoff et al. [**37**] also using Classification and Regression Trees (CART) try to discover lower rainfall, no vector control, higher minimum temperature and houses near breeding sites were associated by order of importance to higher anopheles density to rank malaria risk transmission in Burundi highlands.

From the above work and to the knowledge of the researcher, no tangible work done in Ethiopia to apply knowledge discovery techniques by using local real records to show the relative importance and predictive value of different factors for addressing malaria burden. Even though WHO utilized the data by categorizing malaria burden into inpatient or outpatient, age (less than 5 and greater than 5), malaria type (VIVAX and FALCIPARUM), pregnancy and severe anemia (cases and deaths) are essential but the researcher believe this record utilized for other purposes beside counting and reporting.

Hence, no tangible evidence (patterns), trends and justification to predicate whether malaria deaths/cases are happened or not, high or low using detailed conceptual malaria risk factor model. For example: - sometimes age by itself may not speed up the occurrence of malaria deaths and cases rather their synergic effect with temperature, altitude, type of visit and malaria type increase/decrease the cases or deaths.

All in all, this research is to determine the hierarchical importance of different risk factors and their patterns on malaria death occurrence /cases identification using classification and association rule discovery data mining techniques. Also the paper updates current knowledge and builds a detailed conceptual model for malaria risk factors. To this end, this research will attempt to answer the following questions.

- What are the most hierarchical risk factors in determining malaria occurrence of deaths and type of cases identification that are used as a supporting tool for malaria prevention and control intervention program?
- To what extent data mining techniques help us to identify similar characteristics of malaria risk pattern/factors and the most important determinant factor?
- How to segment/classify the occurrence of malaria type of case identification and occurrence of death (probable or not probable) prediction based on the identified dataset that gives support the current malaria prevention and control intervention program and identify that best fit to the current service (profitable or not) capacities of malaria prevention and intervention programs?

1.3. Objective

The general and specific objectives that this research attempts to achieve are listed as follows:

1.3.1. General Objective

The general objective of this research is to investigate the potential applicability of data mining technologies to discover hidden knowledge in malaria data and develop a model that can help in predicting trend of malaria causes and deaths in order to address malaria health related issue. It focuses on determining the most hierarchical risk factors and their patterns that affect malaria occurrence of deaths and type of cases identification.

1.3.2. Specific Objectives

In order to achieve the above general objective properly, the research attempt to achieve the following specific objectives.

- To review different literatures that can support the study. These are used to:
  - Indicate how malaria is a critical public health issue.

- Show how data mining technologies easily solve and give support to malaria intervention and control program specially by build a model/rules

- To select data mining techniques suitable for investigation and show the characteristics of disease pattern which is important for the prevention and control interventions of malaria in the country.
- To pre-process the data set for training and testing in order to adjust inconsistent data encoding, accounting for missing values, and deriving missing fields from existing ones.
- To build a prototype (predictive model) using WEKA software so as to uncover the knowledge in the training datasets.
- To evaluate the performance of the predictive model built using test dataset and propose further research direction.

1.4. Scope and Limitation of the Study

This research appraises the potential applicability of data mining technology in addressing public health issues in Ethiopia specifically in the case of malaria. The scope of the research mainly focuses on monthly routine case based reports collected from zonal health facility level (i.e. Zonal WHO sites) in Ethiopia. This experimental research undertaking is strictly limited to appraising the potential applicability of data mining technology using classification (J48, MLP and JRip) and association pattern discovery (Apriori) to support malaria prevention and control intervention program in Ethiopia.

This research model limited on the data collected from zonal health post level and it doesn't contain the community level dataset. The dataset also didn't include malaria transmission factors like economic, genetic and social factors as well as list of outpatient death. And also confidentially it doesn't assure or classify the generated model or rules in terms of expected, unrelated and surprising based on user expectation.

1.5. Research Methodology

For this specific research CRISP-DM is adopted. Kurgan and Musile [**46**] survey on knowledge discovery and data mining process models clearly figure out CRISP-DM is the most suitable for novice data miners and especially miners working on industrial projects. This is due to its easiness to read documentation and intuitive,

industry-applications-focused description. It is a very successful and extensively applied model from the bases of grounding its development on practical, industrial and real-world experience.

CRISP methodology is an iterative, adaptive process and involves six phases. That is, the next phase in the sequence often depends on the outcomes associated with the preceding phase. The most significant dependencies between phases are indicated by the arrows. For example, suppose that we are in the modeling phase. Depending on the behavior and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase [**25**]. These are business understanding, data understanding, data preparation and pre-processing, model building, evaluation and deployment steps [**25**]. Hereunder, the researcher tries to briefly discuss methods used following CRIP-DM methodology to achieve the objective of the present research.

1.5.1 Business Understanding

Business understanding enables to get the necessary information as to how the existing system works, which will help to see the applicability of data mining techniques in discovering the most important patterns in malaria data.

In this study both primary data collection (such as observation and interview) and secondary data collection technique such as document analysis (analyzing policies, strategies) are followed to understand malaria related issues, activities and policies in Ethiopia.

1.5.2 Data Collection and Understanding

The researcher surveys there are huge amount of malaria data in Ethiopia that contain detailed information colleted from zonal health facility service using WHO integrated Disease Surveillance form. As a result, there is a need to apply powerful analysis tool to data since classical statistics and database reports may not answer this question of discovering hidden patterns from large volume of data because mostly they discover descriptive summary.

To understand the data we used observation, interviewing with experts as well as the data managers, reviewing documents, reports and literatures done. For example we integrate decisive attributes from Ethiopian National Metrological Agency and Mapping Agency by reviewing the finding on critical factor on malaria transmission

such as temperature, rainfall and altitude [**26, 27, 28, 29, 31, 33**] and discussing with experts.

1.5.3 Data Preparation

To maintain the data quality data completeness, clarity and consistency as well as enhanced participation of data manager as well as experts involved. Also scientific methods used to handle missing, noisy and outlier values, and integration and transformation in order to prepare the data for analysis and build suitable format for Weka data mining tool.

1.5.3.1 Selection

In this phase, Ms- Excel is used for data preparation, pre-processing and summary of statistical analysis tasks for its filtering capability of attributes with different values and SPPS also used for Pre-processing tasks in particular for detecting and statistical summary measures due to its visualization and analysis power.

1.5.3.2 Cleaning, Smoothing and Visualization

Before data pre-processing, an attempt has been made on statistical summary of the data to visualize correct values, missing values and noisy values using Ms-Excel on data sources.

Missing value handled using a global constant (replace all missing attributes values by the same constant) and using the most probable value to fill in the missing value in Weka (determined with regression, inference-based tools using Bayesian formalism or decision tree induction). The required data visualized to assess its representatives and verify data quality.

1.5.3.3 Dataset Construction and Integration

Once we smooth the data we construct target dataset from the initial dataset by extracting and identifying the entity or variables that are important for data integration between flat files as well as data sources. After smoothing the data, data integration made to avoid redundancy, duplication (where there are two or more identical tuples for a given data entry case), inconsistencies, detection and resolution of data value conflicts as well as denormalized table to improve performance by avoiding joins [**11**].

1.5.3.4 Data Transformation and Formatting

Finally, the data transformed into a form that are appropriate for mining with the involvement of smoothing (remove noise), aggregation (summary operation), generalization (low level data replaced into high level concepts), normalization (attribute data are scaled so as to fall within a small specified range) and attribute construction (attribute constructed and added)

1.5.4 Modeling

This study is intended to use J48, JRip and MLP classification techniques with the support of association (Apriori) techniques in WHO malaria data by integrating with national metrological data and Ethiopian mapping agency.

This phase used to build a model (rules) that envisage the occurrence of malaria death, and type of case as well as to discover pattern based on the identified dataset that gives support the current malaria prevention and control intervention program.

To this end, the data mining techniques such as classification algorithms (J48, JRip, MLP) and association rule discovery (Apriori) are selected and calibrate model settings to optimize results. J48 algorithm enables to generate outputs both in tree form and rule sets) and Rule Induction using JRip algorithm (JRip has both the ability and potential to produce accurate but readable rules [**40**]) for the predictive modeling, on the other hand Apriori algorithm for mining the rule induction part. In other words, J48 graphically displays the classification process of a given input for given output class labels **[26, 49, 50**] and rule sets are generally easier to understand since each rule describes a specific context associated with a class and also shows the hierarchy of the determinant factors or attributes [**10, 11, 21, 49, 50]**. WEKA 3.7.3 software is used for building and evaluating models and analysis of the models to provide a uniform interface to many different learning algorithms, along with methods for pre- and post processing and for evaluating the result of learning schemes on any given dataset. It's selected since it supports the whole process of experimental data mining such as preparation of input data, statistical evaluation of learning schemes, visualization of input data and the result of learning and used for education, research and applications.

1.6.5 Evaluation

Different paradigms applied for evaluation to measure interestingness of the rule which provides an overall measure of pattern values combining novelty, usefulness, and simplicity to achieve a predefined goal. The evaluation of results is executed most commonly by combining both an expert and testing tools approaches.

The researcher use different multitude of measurement concentrating on the evaluation of rules such as accuracy, support level, confidence level and complexity with a 10-fold cross validation(the accuracy estimate is the overall number of correct classification from the k iteration divided by the total number of samples, which is k). The first phase of evaluation is analyzing the TP Rate (True Positive Rate), FP Rate (False Positive Rate), TN Rate (True Negative Rate), FN Rate (False Negative Rate), Precision, Recall and Accuracy based on the confusion matrix to establish whether some important facet of the business or research problem has not been accounted for sufficiently and come to a decision regarding use of the data mining results.

Association rules were evaluated in terms of the number of rules and meaning of patterns generated at different minimum support and confidence thresholds for measuring interestingness of the rules. Association was analyzed in terms of different criteria. The criteria include the number of rules generated at different minimum support and confidence thresholds. The minimum support and confidence thresholds varied from 0.1 to 1 and 0.5 to 1 respectively.

Furthermore we investigate the following indicators of the quality of the rule ranking induced by the interestingness measures of the mining algorithm the average rank of the first rule that covers a test instance and the average rank of the first rule that covers and correctly predicts a test instance.

In addition we have two measures by which to evaluate the compactness of an approach the number of mined rules generated by a class association rule miner and the number of rules used for classification. Another important property of an association rule and classification mining algorithm rather than the quality of the rule sets is its time complexity. Therefore we measure the time required for mining and pruning.

These measures are adequate for comparing the whole process of classification and association rules as well as providing the basics for comparing the quality of the mined rule sets of different classification and association rule mining algorithms.

1.6. Significance of the Study

This research can also be used as the corner stone for further studies in the area. It enables us to evaluate the efficiency of data mining techniques in the area of investigating clinical data particularly in the area of malaria. J48 and JRip can predict malaria occurrence of death and case identification as well as general and class association mining to identify factors/patterns. The numeric disease or case value moving from 0 (0%) to 1 (100%) supported the malaria deaths occurrence or case identification being located in alternative risk bands. Categorical risks, such as "probable", "not probable" and so on, based on the data sets, are familiar and easily realized for detecting risk patterns. This might help the people or the health workers be aware, and make more exact calls about the risk predictions for malaria control or intervention. Moreover, the categorical risks enable the use of standard measurement evaluations in order to analyze results from the use of alternative classifiers with these outcomes.

More specifically, this research work can be used to support and strengthen the already implemented malaria prevention and control intervention process by building interesting rules and models because it focuses on developing a malaria prevention and intervention control profile and providing relevant malaria disease pattern information for the specific area which reduces malaria risks by identifying the pattern and cluster group. They also used the research result as a one input for keeping health care malaria risks and show a model for other hot health issues how data mining is important to tackle the problem. In general the health center will get benefit from the project result to keep and improve public health services in Ethiopia.

1.7. Ethical Consideration

The study carried out by considering ethical clearance. Data will be collected after getting permission from World Health Organization Addis Ababa, Ethiopia and informed verbal consent will be obtained. The purpose of the study and privacy during data collection were insured and the dataset has no personal identifier.

1.8. Dissemination of the Research Finding

Result of the research will be present through annual students and staff research conference in the Addis Ababa University for academia, national conference of Ethiopian public health association and will be sent to Journals as an article based on their publishing standards of rules and it will also be communicated to display on Addis Ababa University official web site for authorized individuals for reference purpose for those who are interested in the area as well as putting the hard copy in the libraries for the concerned organization. It also communicated to the organization, Federal Ministry of health, WHO, and responsible bodies.

1.10. Organization of the Thesis

This thesis is organized into seven chapters. The first chapter mainly focuses on an introduction of the research, which contains background of the research work, statement of the problem, objective and methodology of the research and. The second chapter discuss on detail of literature review about knowledge discovery technology, methods/techniques used, and its application in the health care sector and related works. The third chapter discuss mainly on techniques adopted for this research. The fourth chapter discuss the exiting system structure (integrate diseases surveillance system), a means to collect malaria information and epidemiology of malaria. Also, the chapter devoted on data preparation activities and model selection, model implementation and descriptions of Experimentations. The fifth chapter deals with model building using J48, MLP and JRip scenarios respectively. As well as experiment and analysis of classification models using various criteria. Chapter sixth shows the experiments and analysis of association models. Chapter Seven deals with the discussion of the results, final concluding remarks and recommendations forwarded based on the research findings.

CHAPTER TWO
KNOWLEDGE DISCOVERY AND ITS APPLICATION IN HEALTH CARE

2.1. Malaria Situation in Ethiopia

Malaria transmission season in Ethiopia runs from September to December, following the major rainy season from June to August, with a minor transmission season from April to May in areas that receive rains during the short rainy season from February to March. Localized or widespread malaria epidemics can occur during the transmission season. The widespread epidemics have a cyclical pattern of 5 to 8 years that follows major climatic changes [**7**].

The Federal Democratic Republic of Ethiopian Ministry of Health Policy Plan and Finance General Directorate **[13]** published report on health and health related indicator expansion of health extension program on top leading causes of outpatient visits and admission in 2010. The report shows that malaria is still severe case and a headache for the country. Table 2.1, 2.2 and 2.3 shows that detail rank and case of malaria in Ethiopia, annual outpatient and inpatient monthly reportable disease by region in respectively.

Table 2.1 show top ten leading causes of outpatient visits and admission in Ethiopia in the year 2009/2010. Malaria clinical without laboratory confirmation is the first top ranked one causes of outpatient visits with 8.3% and 201, 945 causes and also leading outpatient visits for females with 7.8% and 90, 712 causes. It is also the second top rank outpatients visits for children <5 years with 9.9% and 42, 994 causes and the top four rank leading causes of admission with 4.8% and 5, 017 causes as well as the leading causes of females admission with 4.4% and children <5 years with 6.1%. Of the top ten leading causes of outpatient visits malaria confirmed with species other than P. FALCIPARUM put in top rank with 3.7%, rank nine for females with 3.4% and top ten for children <5 years with 3.4%. Among the leading causes of admissions malaria confirmed with P. FALCIPARUM take rank five with 3.7%, rank six for females with 3.2% and rank four for children < 5 years. In general the table shows malaria is still severe and headache for the country because it is one of the top ten leading causes of outpatient visits and admission.

Table 2.1 top 10 leading causes of outpatient visits and admission in Ethiopia by the year 2001 E.C (2009/2010) [**13**]

Top 10 Leading Causes of Outpatient Visits			
Rank	Diagnosis	Cases	%
1	Malaria (Clinical With out Laboratory Confirmation)	201, 945	8.3
7	Malaria (Confirmed with species other than P. FALCIPARUM)	90, 249	3.7
Top 10 Leading Causes of Outpatients Visits for Female			
Rank	Diagnosis	Cases	%
1	Malaria (Clinical With out Laboratory Confirmation)	90, 712	7.8
9	Malaria (Confirmed with species other than P. FALCIPARUM)	39, 463	3.4
Top 10 Leading Causes of Outpatient Visits for children <5 years (*)			
Rank	Diagnosis	Cases	%
2	Malaria (Clinical with out laboratory confirmation)	42, 994	9.9
10	Malaria (Confirmed with species other than P. FALCIPARUM)	16, 140	3.4
Top 10 Leading Causes of Admission			
Rank	Diagnosis	Cases	%
4	Malaria (Clinical with out laboratory confirmation)	5, 017	4.8
5	Malaria (Confirmed with P. FACLIPARUM)	3,821	3.7
Top 10 Leading Causes of Admission for Females			
Rank	Diagnosis	Cases	%
4	Malaria (Clinical with out laboratory confirmation)	2,177	4.4
6	Malaria (Confirmed with P. FACLIPARUM)	1, 573	3.2
Top 10 Leading Causes of Admission for Children < 5 years (**)			
Rank	Diagnosis	Cases	%
3	Malaria (Clinical with out laboratory confirmation)	673	6.1
4	Malaria (Confirmed with P. FACLIPARUM)	669	6.1

** This shows the report doesn't include SNNPR, Somali and Addis Ababa except Federal Hospital Morbidity report & ** This shows the report doesn't include SNNPR, Afar, Somali and Addis Ababa except Federal Hospital Morbidity report*

*Table 2.2 annual outpatient monthly reportable malaria diseases by region in the year 2001 E.C (2009/2010) [**13**]*

Region	Malaria <5 Year		Malaria > 5 year		Malaria in Pregnancy
	Total	Lab Confirmed	Total	Lab Confirmed	
Tigray	21, 733	5, 125	98, 946	17, 685	847
Afar	6, 669	2, 556	26, 472	6, 205	413
Amhara	38, 439	5, 580	136, 930	22, 582	1, 747
Oromia	50, 926	10, 554	143, 019	25, 975	2, 994
Somali	NR	NR	NR	NR	NR
Ben-Gumuz	12, 836	895	23, 062	2, 105	452
SNNPR	138, 689	36, 244	390, 368	105, 871	7, 532
Gambella	3, 186	846	7, 365	2, 259	922
Harar	166	27	664	300	4
Addis Ababa	280	116	2, 018	567	3
Dire Dawa	98	2	291	5	0
National	275, 022	61, 945	829, 135	183, 554	14, 864

Table 2.2 clearly shows that annual outpatient monthly reportable malaria diseases by region in the year 2009/2010. SNNPR, Oromia, Amhara and Tigray are the leading region who report malaria causes and lab confirmed causes greater than 5 years and less than 5 years respectively. Oromia, Amhara and SNNPR are leading causes of malaria in pregnancy.

Table 2.3 annual inpatient malaria monthly reportable disease by region in Ethiopia in the year 2001 E.C (2009/2010) **[13]**

Region	Malaria < 5 Years		Malaria >5 Years		Malaria In Pregnancy	
	Cases	Deaths	Cases	Deaths	Cases	Deaths
Tigary	258	10	1, 573	103	55	0
Afar	37	2	140	2	10	0
Amhara	190	7	696	15	24	0
Oromia	1,049	39	1,925	52	96	5
Somali	NR	NR	NR	NR	NR	NR
Ben-Gum	159	0	200	3	0	0
SNNPR	3, 347	92	10, 009	247	373	5
Gambella	112	5	313	6	13	0
Harar	17	0	36	0	0	0
Addis Ababa	8	0	52	0	3	0
Dire Dawa	1	0	8	2	0	0
National	5,178	155	14, 952	430	574	10

Table 2.3 clearly shows that annual inpatient monthly reportable malaria diseases by region in the year 2009/2010. SNNPR, Oromia and Tigray are the leading region who report malaria causes and deaths above 5 years and below 5 years respectively. SNNPR, Oromia and Tigray are leading causes of malaria in pregnancy and Oromia and SNNPR region report malaria in pregnancy deaths.

2.1.1. Prevalence of Malaria in Ethiopia

Malaria is seasonal in most parts of Ethiopia, the transmission patterns and intensity vary greatly due to the large diversity in altitude, rainfall, and population movement; areas below 2,000 meters are considered to be malarious (or potentially malarious). Those areas are approximately 68% of the Ethiopian population and cover almost 75% of the country's landmass [**5**]. The malaria distribution and seasonality of malaria in Ethiopia by altitude listed in the Table 2.4.

*Table 2.4 The Distribution and Seasonality of Malaria in Ethiopia [**6**].*

Altitude	Distribution and seasonality of malaria
>2500m	Malaria free highlands
>=2000 and <=2500m	Highlands affected by occasional epidemics
<1500m with rainfall >1000mm	Malarious lowlands with intense transmission

Different researches were conducted on malaria to indicate the malaria death and prevalence rate in Ethiopia. Among this the most popular one was Ethiopian malaria indicator survey in 2007. The research result shows the malaria prevalence rate by species and location as well as by severe anemia presented in table 2.5

*Table 2.5 Parasite prevalence rates, by species and location, 2007[**5**]*

	Parasite Prevalence (%)		
	P. falciparum	*P.vivax*	Total
Nationwide	0.5	0.2	0.7
Malarious Areas (below 2, 000m)	0.7	0.3	0.9

Table 2.5 clearly indicates that by microscopy, parasite prevalence in all ages was 0.7%, with 76% of infections being P. faclciparum. The survey shows that severe

anemia prevalence among children under age five years with severe anemia were 5.5% nation wide and 6.6% in malarious areas (below 2,000m) [**5**].

Schunk et al [**18**] conduct a research on high prevalence of drug resistance mutations in plasmodium falciparum and plasmodium vivax in southern Ethiopia. They conduct on 100 patients with uncomplicated malaria from Dilla, southern Ethiopia, P. falciparum DHFR and DHPS mutations as well as P. vivax dhfr polymorphisms associated with resistance to SP and P. falciparum pfcrt and pfmdr1 mutations conferring CQ resistance were assessed. Result they obtained shows that P. falciparum and P. vivax were observed in 69% and 31% of the patients, respectively. Pfdhfr triple mutations and pfdhfr/pfdhps quintuple mutations occurred in 87% and 86% of P. falciparum isolates, respectively. Pfcrt T76 was seen in all and pfmdr1 Y86 in 81% of P. falciparum. The P. vivax dhfr core mutations N117 and R58 were present in 94% and 74%, respectively.

Karunamoorthiab and Bekelea [**19**] also conduct a research on Prevalence of malaria from peripheral blood smears examination: A 1-year retrospective study from the Serbo Health Center, Kersa Woreda, Ethiopia between July 2007 and June 2008. Of the total 6863 smears, 3009 were found to be positive and contribute 43.8% of diagnostic yield. Plasmodium falciparum constituted the most predominant [64.6%], while Plasmodium vivax confirmed with 34.9% cases. Among patients who underwent diagnostic testing and treatment for malaria, males were more prone to have a positive malaria smear than females.

2.2. Overview of Data Mining

The need to understand large, complex, information-rich data sets is common in virtually all fields such as business, health, science, and engineering [**10**]. For example, in the business world, corporate and customer data are becoming recognized as a strategic asset. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer-based methodology, including new techniques, for discovering knowledge from data is called data mining [**20**]. Different scholars give different definitions of data mining as listed in the table 2.6.

Table 2.6: Definitions of Data Mining

Scholars	Definition
Han Jiawei and Kamber Micheline [**11**]	Data mining refers to extracting knowledge from large amount of data. It is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses and other information repositories.
Mehmed Kantardzic [**20**]	Data mining is an iterative process of discovering an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.
Witten I.H. and Frank E. [**14**]	Data mining is about solving problems by analyzing and discovering patterns in data already present in databases. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The process must be automatic or (more usually) semiautomatic and the data is invariably present in substantial quantities.
David Hand, Heikki Mannila and Padhraic Smyth [**21**]	Data mining is the analysis of (often large) observational data sets to discover unsuspected relationships between attributes and to summarize the data in novel ways that are both understandable and useful to the data owner.
Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees and Alessandro Zanasi [**22**]	Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of knowledge extraction from large data bases.
S. Sumathi and S.N. Sivanandam [**23**]	Data mining is concerned with finding hidden relationships present in business data to allow businesses organization to make predictions for future use. It is the process of data-driven extraction of not so obvious but useful information from large databases. Data mining has emerged as a key technology for business intelligence.

2.3. Data Mining Process

Discovering knowledge in data presents data mining as a well-structured standard process, intimately connected with managers, decision makers, and those involved in deploying the results [**8**].

Many people treat data mining as synonym for another popularly used term, knowledge discovery from data or KDD [**11**]. Alternatively others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process is depicted in figure 2.1 and consists of an iterative sequence of the steps of Data cleaning, Data integration, Data selection, Data transformation, Data mining and Pattern evaluation.

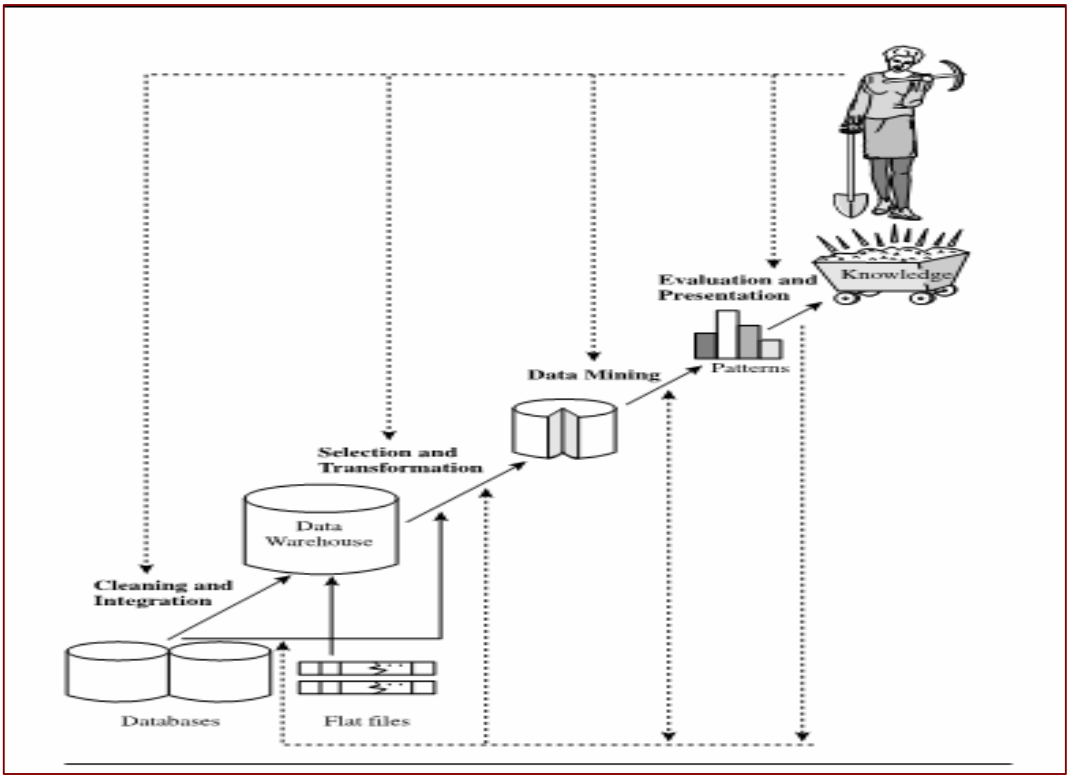


*Figure 2.1: Data mining as step in the process of knowledge discovery [**11**]*

Maytt [**11**] also noted any exploratory data analysis/data mining project should include Problem definition, Data preparation, Implementation of analysis and Deployment of results.

In order to make sense of the huge amount of data, Azevedo and Manuel [**48**] survey on referred methodologies or process indicates KDD (Knowledge Discovery in Databases), SEMMA (Sample Explore Modify Model Assess) and CRISP-DM

(Cross-Industry Standard Process for DM) are most cited and commonly used. They noted SEMMA and CRISP-DM can be viewed as an implementation of the KDD process described by Fayyad et al [**48**] and CRISP-DM is more complete than SEMMA. Kurgan and Musilek [**46**], also survey on the five major models and has been identified different steps that need to be carried out in order to explore different interesting rules.

As Kurgan and Musile [**46**] noted the common steps among the DM models are domain understanding, data preparation, data mining and evaluation. The main difference listed as follows

1. Fayyad's nine-step model performs activities related to DM task and algorithm relatively late in the process. In this model prepared data may not be suitable for the tool of choice, and thus a loop back step may be required. While the other models perform before pre processing (i.e. the data are correctly prepared for the DM step without the need to repeat some of the earlier steps [**46**]).
2. Cabena's model omits the Data Understanding step and this part is filled by adding Data Audit between data preparation and DM by Hirji [**46**], who used this model in a business project and very similar to Cios and CRISP-DM model.
3. The eight-step model by Anand & Buchner provides a very detailed breakdown of steps in the early phases of the process but it does not include activities necessary for putting the discovered knowledge to work.
4. CRISP-DM developed based on several companies' significant industrial input and involvement together with academic and governmental support. It is a very mature model that has been thoroughly documented and tested in many applications.
5. The six-step model by Cios emphasizes academic aspects of the process. It is also the only model that provides detailed guidelines concerning possible loops, rather than just mentioning their presence and draws significantly from the CRISP-DM model.

The Cross-Industry Standard Process for Data Mining (CRISP–DM) **[25]** was developed in 1996 by analysts representing DaimlerChrysler, SPSS and NCR. CRISP provides anon proprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit.

According to CRISP–DM, a given data mining project has a life cycle consisting of six phases. The iterative nature of CRISP is symbolized by the outer circle in Figure 3.1 [**25**]. Often, the solution to a particular business or research problem leads to further questions of interest, which may then be attacked using the same general process as before.

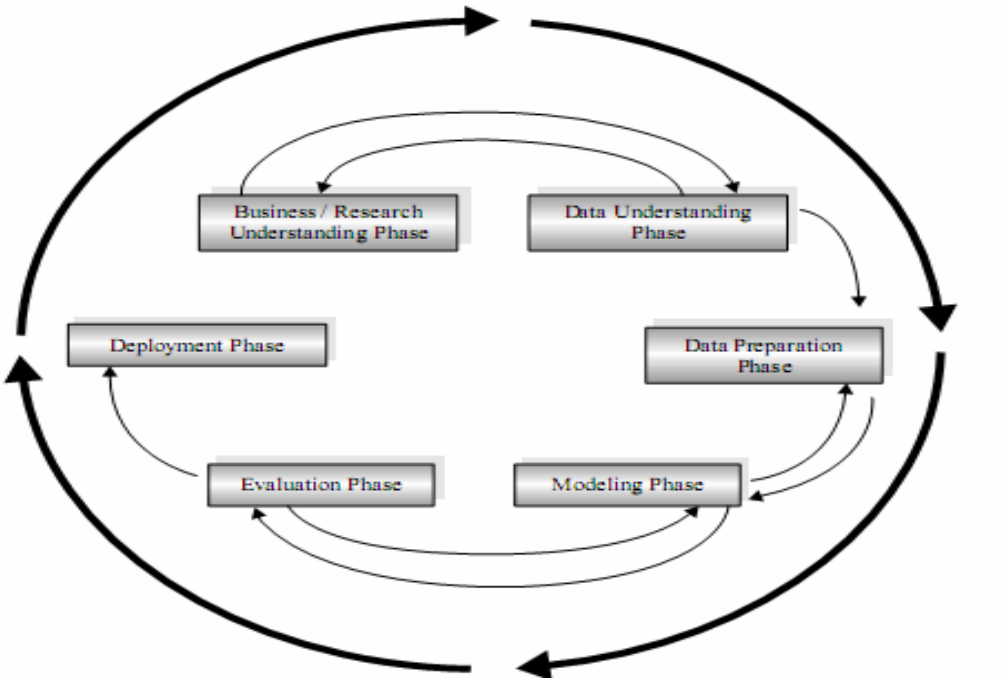


*Fig 2.2: CRISP-DM [**25**]*

CRISP methodology is an iterative, adaptive process and involves Six Phases. These are business understanding, data understanding, data preparation and pre-processing, model building, evaluation and deployment steps **[25]**. Here under, the researcher tries to briefly discuss methods used following CRIP-DM methodology to achieve the objective of the present research.

Business Understanding Phase **i**n the CRISP–DM standard process which may also be termed the research understanding phase [**25**]. It is a decisive DM course of action to understand the research area as well as problem domain.

Data Understanding Phase**,** the required data is collected and visualized to assess its representatives and verify data quality.

Data Preparation Phase data selection, pre-processing, integration and transformation implemented to formulate for analysis and build suitable format. Smoothing is used to avoid inaccurate data entry or updating. Missing values handle using a global constant (replace all missing attributes values by the same constant) and using the most probable value to fill in the missing value (determined with regression, inference-based tools using a Bayesian formalism or decision tree induction [**11**]. Han and Kamber [11] stated outlier handled using clustering method and noisy values (random errors) handled by smoothing the data through binning and consulting its "neighborhood". A common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5*IQR above the third quartile (Q3) or below the first quartile (Q1). In other words it is to mean that the values outside the limits Q3+ (1.5*IQR) and Q1-(1.5 *IRQ) will be considered outlier values.

After smoothing the data, data integration made to avoid redundancy, duplication (where there are two or more identical tuples for a given data entry case), inconsistencies, detection and resolution of data value conflicts as well as denormalized table to improve performance by avoiding joins [**11**].

Finally, the data transformed into a form that are appropriate for mining with the involvement of smoothing (remove noise), aggregation (summary operation), generalization (low level data replaced into high level concepts), normalization (attribute data are scaled so as to fall within a small specified range) and attribute construction (attribute constructed and added).

Modeling phase used to build a model (rules) that envisage the occurrence of malaria death, and type of case as well as to discover pattern based on the identified dataset that gives support the current malaria prevention and control intervention program.

Evaluation phase mainly measures the performance of one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field, determine whether the model in fact achieves the objectives set. The assessment of knowledge generated during the knowledge discovery process is usually approached a domain expert manually analyzes the generated knowledge and judges its usefulness and interestingness according to their own knowledge and established project goals and the other approach performs a more formal evaluation,

usually involving statistical tests via cross-validation or more advanced test schemas [**46**].

Different paradigms applied for evaluation to measure interestingness of the rule which provides an overall measure of pattern values combining novelty, usefulness, and simplicity to achieve a predefined goal. The evaluation of results is executed most commonly by combining both an expert and testing tools approaches. Usually a large number of interesting patterns are mined [**46**], and the formal evaluation is used to sort out all irrelevant or obvious cases before a human expert performs assessment. And final phase involves generation of a report.

2.4. Data Mining Functionalities and Techniques

A database is a store of information but more important is the information which can be inferred from it. In order to do this a wide variety of data-mining methods or techniques should be used. There is no particular rule that would tell you when to choose a particular technique over another one. Sometimes those decisions are made relatively arbitrary based on the availability of data mining analysts who are most experienced in one technique over another. These techniques can be used for either discovering new information within large databases or for building predictive models [**12**].

Kantardzic [**20**] explained the primary goals of data mining are prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. As Kantardzic [**20**] noted on the predictive end of the spectrum, the goal of data mining is to create a model, expressed as an executable code, which can be used to perform classification, prediction or estimation. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. According to Kantardzic [**20**], the goal of descriptive end of the spectrum, is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for particular data-mining applications can vary considerably [**20**].

In general data mining methods or techniques may be classified by the function they perform or according to the class of application they can be used in. The most popular

data mining tasks are classification, clustering and association rule discovery [**10, 11, 21, 22, 23, 24**].

2.4.1. Classification and Prediction

Classification is the process of learning a function that maps (classifies) a data item into one of several predefined classes [**14**]. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from a historical database. The main objective of classification is to identify the characteristic that indicate the group to which each case belong. This pattern can be used both to understand the existing data and to predict how new instances will behave. That is, the system takes a case or records with certain known attribute values and able to predict what class this case belongs to. Prediction can be viewed as the construction and use of a model to asses the class of unlabeled sample is likely to have. That means, it predict unknown or missing class value [**14**].

There are various classification techniques among which the common once are the following.

2.4.1.1. Decision Trees

A decision tree is defined as hierarchical knowledge representation techniques and it classifies examples/records to a finite number of classes. The nodes are labeled with attribute names, the edges are labeled with possible values for this attribute and the leaves labeled with different classes [**14**].Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes.

Because of their tree structure and ability to easily generate rules, decision trees are the favored technique for building understandable models. Decision trees are also a natural choice when the goal is to generate rules that can be easily understood, explained, and translated into standard query language or natural language [**14**].

Decision trees are a wonderfully versatile tool for data mining. There are two main types of decision trees [**12**]:

1. Classification tree- takes categorical values and label records and assigns them to the proper class. The classification tree reports the class probability, which is the confidence that a records is in a given class.

2. Regression tree- estimates the value of a target variable that takes on numeric values.

All decision tree construction methods are based on the principle of recursively partitioning the dataset until homogeneity is achieved. Various listed decision tree algorithms are available such as CHAID (Chi-Square Automatic Interaction Detector), C4.5/C5.0, CART (Classification and Regression Trees), ID3 and many others differ from one another in the number of splits allowed at each level of tree, how those splits are chosen when the tree is built [**14**].

2.4.1.2. Rule Induction

Rule induction (sometimes called rule learner) is one of the major forms of data mining techniques and is the most common form of knowledge discovery learning systems. It is also perhaps the form of data mining that most closely resembles the process that most people think about when they think "mining" for gold through a vast database. The gold in this case would be a rule that is interesting - that tells you something about your database that you didn't already know and probably weren't able to explicitly articulate [**14**].

Rule induction systems are highly automated and are probably the best of data mining techniques for exposing all possible predictive patterns in a database. They can be used in prediction problems but the algorithms for combining evidence from a variety of rules come from practical experience [**14**].

2.4.1.3. Neural Networks

Neural networks use a set of processing elements (or nodes) analogous to neurons in the brain [**14**]. These processing elements are interconnected in a network that can then identify patterns in data once it is exposed to the data, i.e. the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs that simply follow instructions in a fixed sequential order.

A neural network consists of a layered, feed forward, completely connected network of nodes [**24**]. The feed forward nature of the network restricts the network to a single direction of flow and does not allow looping or cycling.

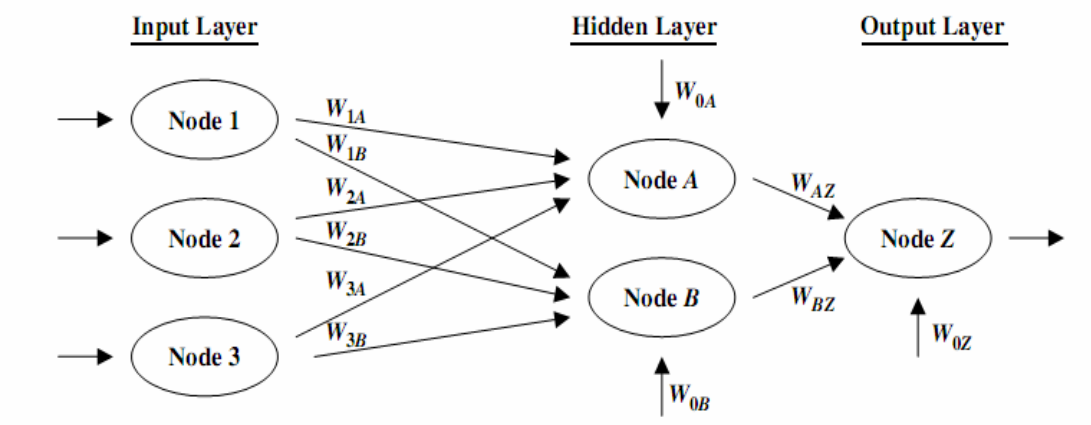


*Figure2.3. simple neural network [**17**]*

The neural network is composed of two or more layers, although most networks consist of three layers: an input layer, a hidden layer, and an output layer just like listed in the figure 2.3. There may be more than one hidden layer, although most networks contain only one, which is sufficient for most purposes. The neural network is completely connected, meaning that every node in a given layer is connected to every node in the next layer, although not to other nodes in the same layer. Each connection between nodes has a weight (e.g., W) associated with it. At initialization, these weights are randomly assigned to values between zero and 1. The number of input nodes usually depends on the number and type of attributes in the data set. One may have more than one node in the output layer, depending on the particular classification task at hand [**24**].

The broad applicability of neural networks to real world business problems have already been successfully applied in many industries. Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting needs including sales forecasting, industrial process control, customer research, data validation, risk management etc [**14**].

In conclusion, survey in existing algorithm comparative studies shows that decision tree algorithms run significantly faster during training where us the neural network always perform better at classifying novel examples in the presence of incomplete and noisy data as well as It outperforms in classifying real world dataset because real world datasets are often not linearly separable as well as slow to train [**26**].

2.4.2. Association Rule Discovery

Association analysis is the discovery of association rules showing attribute value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis. Given a collection of items and a set of records, each of which contain some number of items from the given collection, an association function is an operation against this set of records which return affinities or patterns that exist among the collection of items [**11**].

The association task for data mining is the job of finding which attributes "go together". Association rules take the form of "If antecedent, then consequent," along with a measure of the support and confidence associated with the rule. For example, in a AMHARA region AWI zone there may occurs 100 malaria case and death on September, 2010, among these there are 50 malaria in pregnancy inpatient cases, and of the 50 pregnancy inpatient cases 20 dies. Thus, the association rule would be: "If there is malaria in pregnancy inpatient cases, then there is a possibility to die," with a support of 20 / 100 =20% and a confidence of 20 / 50 =40%) [**24**].

Association rule mining finds interesting association or correlation relationship among a large set of data items. The discovery of interesting association relationships among huge amounts of business transaction records can help in many business decision making processes, such as catalogue design and cross marketing. A typical example of association rule mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in there shopping baskets [**14**].

There are various association rule discovery techniques among which the common once are the following.

2.4.2.1. Apriori Algorithm

Apriori algorithm is an influential algorithm for mining frequent item sets for Boolean association rule and takes advantage of structure within the rules themselves to reduce the search problem to a more manageable size [**11**, **24, 27**].

In this research Apriori selected/implemented since it's one of the most popular data mining approaches to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or

equal to a user specified minimum support) is not trivial because of its combinatorial explosion [**51**]. The detail of the techniques listed in chapter three.

2.4.2.2. FP-Growth Algorithm

As shown above the main bottleneck of the a priori like methods are at the candidate set generation and test. This problem was dealt with by introducing a novel, compact data structure, called frequent pattern tree, or FP-tree then based on this structure an FP-tree-based pattern fragment growth method was developed, FP-growth. A frequent pattern tree is a tree structure defined below [**27**].

1. It consists of one root labeled as "root", a set of item prefix sub-trees as the children of the root, and a frequent-item header table.
2. Each node in the item prefix sub-tree consists of three fields: item-name, count, and node-link, where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node-link links to the next node in the FP-tree carrying the same item-name, or null if there is none.
3. Each entry in the frequent-item header table consists of two fields, (1) item-name and (2) head of node-link, which points to the first node in the FP-tree carrying the item-name.

2.4.3. Clustering

Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A cluster is a collection of records that are similar to one another, and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering task does not try to classify, estimate, or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized and the similarity to records outside the cluster is minimize [**24**].

Clustering analyze data objects with out consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that

objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [**11**].

Clustering activity involves the pattern representation (including feature extraction and/or selection), definition of a pattern proximity measure appropriate to the data domain, clustering or grouping as represented graphically in the figure 2.3. If needed, it contains data abstraction (extracting a simple and compact representation of a data set) and assessment of output [**28**].

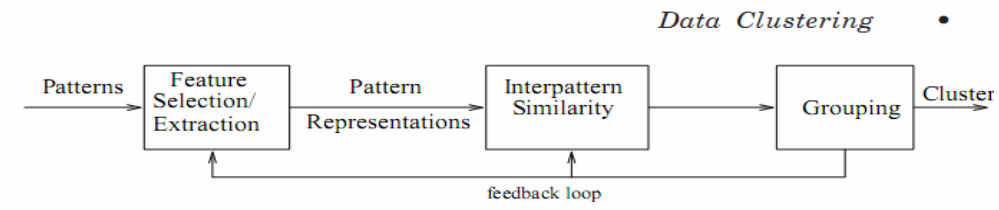


Figure 2.4: Stages in Clustering ***[28]***

There are various clustering techniques as shown in the figure 2.5 among which the common, simple to implement and computationally attractive because of its linear time complexity are K-Means clustering [**28**].

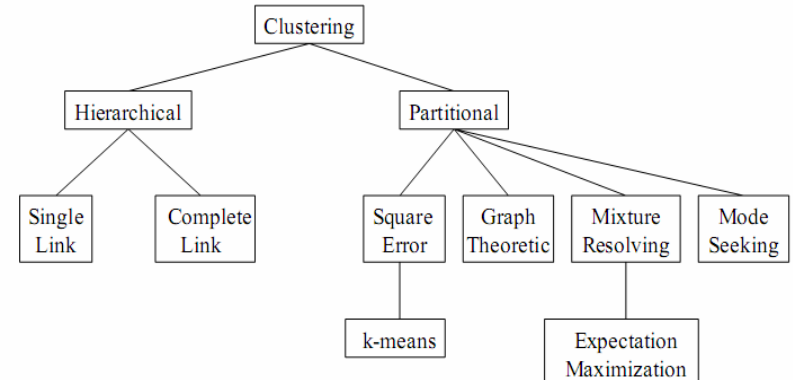


Figure 2.5: Type of Clustering ***[28]***

2.4.3.1. K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The algorithm is composed of the following steps [**24, 28**]:

1. *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*
2. *Assign each object to the group that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

2.5. Review of Data Mining Application in Health Domain

The successful application of data mining in highly visible fields like Financial and Marketing Data Analysis, Telecommunications Industry, Retail Industry, Science and Engineering and e-business have led to the popularity of its use in knowledge discovery in databases (KDD) in other industries and sectors. Among these sectors that are just discovering data mining are the fields of medicine and public health.

Myatt **[10]** also said almost every field of study is generating an unprecedented amount of data. Retail companies collect data on every sales transaction, organizations log each click made on their web sites, and biologists generate millions of pieces of information related to genes daily. The volume of data being generated is leading to information overload and the ability to make sense of all this data is becoming increasingly important. It requires an understanding of exploratory data analysis and data mining as well as an appreciation of the subject matter, business processes, software deployment, project management methods, change management issues, and so on **[10]**.

Besides the above facts, the past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and advances in cancer therapies to the identification and study of the human genome. The logic behind investigating the genetic causes of disease is that once the molecular bases of diseases are known, precisely targeted medical interventions for diagnostics, prevention, and treatment of the disease themselves can be developed [**20**].

With the amount of information and issues in the healthcare industry, not to mention the pharmaceutical industry and biomedical research, opportunities for data-mining applications are extremely widespread, and benefits from the results are enormous.

Storing patients' records in electronic format and the development of medical-information systems cause a large amount of clinical data to be available online. Regularities, trends, and surprising events extracted from these data by data-mining methods are important in assisting clinicians to make informed decisions, thereby improving health services [**20**].

Today, the size of the population, the amount of electronic data gathered, along with globalization and the speed of disease outbreaks make it almost impossible to accomplish what the pioneers did. This is where data mining becomes useful to healthcare. It has been slowly but increasingly applied to tackle various problems of knowledge discovery in the health sector.

As a result, there is a need to apply powerful analysis tool like data mining to this malaria data classical statistics and database reports may not answer this question of discovering hidden patterns from large volume of data. Since there are huge amount of malaria data in Ethiopia that contain detailed information related to health facility name, address, inpatient cases and death and out patient cases collected by using WHO integrated Disease Surveillance form from zonal health facility service center.

Currently in the digital world the rapid growth of Healthcare and Biomedical field storing patient records in electronic format and the development in medical-information systems cause a large amount of clinical data to be available online. Regularities and surprising events extracted from these data by data-mining methods are important in assisting clinicians to make informed decisions, thereby improving health services. Data mining has been used in many successful medical applications, including data validation in intensive care, the monitoring of children's growth, analysis of diabetic patient's data, the monitoring of heart-transplant patients [**20**]. There are different researches conducted on application of data mining on health.

Here below, we tries to review some of the related works that are done in healthcare and public health including malaria that are important for this research.

2.5.1. Mining Health Care Data

Obenshain [**29**] conduct a research on “Applications of Data mining healthcare data”.

He compared data mining with traditional statistics. He identified reduction of time and effort on the part of end user, ability to examine multiple areas simultaneously, decrease potential for human error, correct data presentation and accessibility are

some advantages of automated data systems and described prediction, classification, exploration and affinity analysis are main data mining strategies.

Result he obtained shows that Hospital Infection Control, Ranking Hospitals and Identifying High-Risk Patients are successful health areas that data mining applications have been implemented easily.

In his conclusion, he states that automated surveillance systems offer obvious advantages over manual ones and put recommendation on further exploration of data mining for research related to infection control and hospital epidemiology seems in order, especially where the data volume exceeds capabilities of traditional statistical techniques.

Hirano and Tsumoto [**30**] conduct a research on "Temporal Data Mining in Hospital Information Systems: Analysis of Clinical Courses of Chronic Hepatitis".

Methods they were used to find interesting knowledge from temporal data on chronic diseases are based on the combination of advanced sequence comparison techniques and cluster analysis procedure.

Cluster analysis system was introduced for temporal data to apply analysis of platelet (PLT) count data on chronic viral hepatitis patients and produced PLT value-based temporal analysis that are important for finding years for reaching F4 (liver fibrosis stage four), years elapsed between stages, and their relationships with virus types and fibrotic stages.

Result they obtained conveyed that the temporal courses of PLT could be grouped into several patterns exhibiting similar average PLT level and increase/decrease trends, and liver fibrosis might proceed faster in some exacerbating cases. And they recommend, validating the clinical reasonability of the results and usefulness of the system should be tested using other datasets.

2.5.2. Mining Malaria Data

Junior and Duarte [**31**] conduct a research on "Artificial Neural Networks and Bayesian Networks as Supporting Tools for Diagnosis of Asymptomatic Malaria".

Methods they were used for the diagnosis of asymptomatic malaria infection are Artificial Neural Network (ANN) and Bayesian Network (BN) techniques as supporting tools. These techniques are compared with two classical laboratorial tests for diagnosis of malaria: the light microscopy and molecular test were run in 380 individuals from the Brazilian Amazon.

Results for the Artificial Neural Network (ANN) presented 67.5% of sensitivity and 92.5% of specificity, and have correctly diagnosed 80.0% of the 80 individuals separated for testing the method; the Bayesian Network (BN) presented 37.5% of sensitivity and 97.5% of specificity, and has correctly diagnosed 67.0% of the 80 individuals separated for testing the method and the Microscopy Test The microscopy yielded 22.5% of sensitivity and 100% of specificity, thus, correctly diagnosed 61.25% of the 80 individuals separated for testing. They stated that both innovative techniques are able to identify asymptomatically infected individuals with better accuracy than the microscopy test and are potentially useful for helping the diagnosis of asymptomatic malaria.

In their recommendation, they indicate the next step of the research will be the development of a multi platform executable program, allowing the use of the system in mobile phones as well as to try to use different computational model for example, so that a doctor can easily make the diagnosis even in remote areas.

Wangdi et al [**32**], study was carried out retrospectively using the monthly reported malaria cases from the health centres to Vector-borne Disease Control Programme (VDCP) and the meteorological data from Meteorological Unit in endemic districts of Bhutan, Department of Energy and Ministry of Economic Affairs.

Time series analysis was performed on monthly malaria cases, from 1994 to 2008, in seven malaria endemic districts. The time series models derived from a multiplicative seasonal autoregressive integrated moving average (ARIMA) was deployed to identify the best model using data from 1994 to 2006.

Methods they were used to determine predictors of malaria of the subsequent month are ARIMAX Modeling .

Result they obtained shows that the ARIMAX model of monthly cases and climatic factors show considerable variations among the different districts. In general, the mean maximum temperature lagged at one month was a strong positive predictor of an increased malaria cases for four districts. The monthly number of cases of the previous month was also a significant predictor in one district, whereas no variable could predict malaria cases for two districts.

In their conclusion, they stated that the ARIMA models of time-series analysis were useful in forecasting

the number of cases in the endemic areas of Bhutan and could be employed for planning and managing malaria prevention and control programme.

Roca-Feltrer et al [**33**] conduct a research on the age patterns of severe malaria syndromes in sub-Saharan Africa across a range of transmission intensities and seasonality settings.

They tries to indicate understanding of the relationship between transmission intensity, seasonality and the age pattern of malaria is needed to guide appropriate targeting of malaria interventions in different epidemiological setting.

Methods they were used to report the age of paediatric hospital admissions with cerebral malaria, severe malarial anemia, or respiratory distress are a systematic literature review.

Study sites were categorized into a 3 × 2 matrix of P.FALCIPARUM transmission intensity and seasonality. Probability distributions and best fitting models were used to represent graphically the age-pattern of each outcome for each transmission category in the matrix.

Result they obtained shows that there is a shift in the burden of cerebral malaria towards younger age groups was seen with increasing intensity of transmission, but this was not the case for severe malarial anemia or respiratory distress. Sites with ‘no

marked seasonality' showed more evidence of skewed age-patterns compared to areas of 'marked seasonality' for all three severe malaria syndromes.

In their conclusion, they stated that although the peak age of cerebral malaria will increase as transmission intensity decreases in Africa, more than 75% of all paediatric hospital admissions of severe malaria are likely to remain in under five year olds in most epidemiological settings.

Yé et al [**34**] conduct a research on the effect of meteorological factors on clinical malaria risk among children: an assessment using village-based meteorological stations and community-based parasitological survey. In view of the fact that temperature; rainfall and humidity have been widely associated with the dynamics of malaria vector population and spread of the disease using the data collected at the same time and scale.

676 children (6–59 months) were selected randomly from three ecologically different sites (urban and rural). During weekly home visits between December 1, 2003, and November 30, 2004, fieldworkers tested children with fever for clinical malaria. They also collected data on possible confounders monthly. Digital meteorological stations measured ambient temperature, humidity, and rainfall in each site.

Logistic regression was used to estimate the risk of clinical malaria given the previous month's meteorological conditions.

Result they obtain stated that the overall incidence of clinical malaria over the study period was 1.07 episodes per child. Meteorological factors were associated with clinical malaria with mean temperature having the largest effect.

In their conclusion, they stated that temperature was the best predictor for clinical malaria among children under five. A systematic measurement of local temperature through ground stations and integration of such data in the routine health information system could support assessment of malaria transmission risk at the district level for well-targeted control efforts.o

Teklehaimanot et al [**35**] conducts a research on weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia. Patterns of lagged weather effects reflect biological mechanisms.

Aims they were targeted to understand are the reason for variation is crucial to determining specific and important indicators for epidemic prediction.

Methods they were modeled using a robust regression with rainfall, minimum temperature and maximum temperatures variables are using daily average number of cases and group the districts into hot and cold climatic zone in 10 districts of Ethiopia.

Result they obtained shows that in cold districts, rainfall was associated with a delayed increase in malaria cases, while the association in the hot districts occurred at relatively shorter lags. In cold districts, minimum temperature was associated with malaria cases with a delayed effect. In hot districts, the effect of minimum temperature was non-significant at most lags and much of its contribution was relatively immediate.

As their conclusion; states that the interaction between climatic factors and their biological influence on mosquito and parasite life cycle is a key factor in the association between weather and malaria. These factors should be considered in the development of malaria early warning system.

Newman et al [**36**] conduct a research on Burden of Malaria during Pregnancy in Areas of Stable and Unstable Transmission in Ethiopia during a Non epidemic Year.

Methods they were used to identify peripheral malaria parasitemia are cross-sectional studies i.e. 10.4% of women attending antenatal care clinical at 1 stable transmission site and 1.8% of women at 3 unstable sites.

As they identified placental parasitemia was identified more frequently during deliveries at stable site than unstable sites and associated with low birth weight at the stable site and pre-maturity at stable sites and unstable sites and with a 7-fold increased risk of stillbirths at unstable sites.

As they state in their conclusion the effectiveness and efficiency in Ethiopia of standard preventive strategies used in high transmission regions (such as intermittent preventive treatment) may require further evaluation.

Protopopoff et al. [**37**] conduct a research on Ranking Malaria Risk Factors to Guide Malaria Control Efforts in African Highlands for a better understanding of the factors impacting transmission in the highlands is crucial to improve well targeted malaria control strategies.

Methods they were used to build a conceptual model of potential malaria risk factors in the highlands are based on the available literature, classification and regression trees, and an unexploited statistical method to analyze the risk factors in the Burundi highlands.

Results they obtained show that anopheles density was the best predictor for high malaria prevalence. Then lower rainfall, no vector control, higher minimum temperature and houses near breeding sites were associated by order of importance to higher Anopheles density.

As their conclusion, states that In Burundi highlands monitoring Anopheles densities when rainfall is low may be able to predict epidemics. The conceptual model combined with the CART (classification and regression tree) analysis is a decision support tool that could provide an important contribution toward the prevention and control of malaria by identifying major risk factors.

Yeshiwondim et al [**38**] conduct a research on Spatial analysis of malaria incidence at the village level in areas with unstable transmission in Ethiopia to examines the spatial and temporal patterns of malaria transmission at the local level and implements a risk mapping tool to aid in monitoring and disease control activities.

Method they were used to examine the global and local patterns of malaria distribution are using individual-level morbidity data collected from six laboratory and treatment centers between September 2002 and August 2006 in 543 villages in East Shoa, central Ethiopia.

Results they obtained show that statistical analysis of malaria incidence by sex, age, and village through time reveal the presence of significant spatio-temporal variations. Poisson regression analysis shows a decrease in malaria incidence with increasing age. A significant difference in the malaria incidence density ratio (IDRs) is detected in males but not in females. A significant decrease in the malaria IDRs with increasing age is captured by a quadratic model.

As their conclusion, stated that malaria incidence varies according to gender and age, with males age 5 and above showing a statistically higher incidence. Significant local clustering of malaria incidence occurs between pairs of villages within 1–10 km distance lags.

Roca-Feltrer et al [**39**] conduct a research on defining malaria seasonality to aid localized policymaking and targeting of interventions.

Methods they were used a series of systematic literature reviews were undertaken to identify studies reporting on monthly data for full calendar years on clinical malaria, hospital admission with malaria and entomological inoculation rates.

Results they obtained show that monthly data for full calendar years on clinical malaria, all hospital admissions with malaria, and entomological inoculation rates were available. Most sites showed year-round transmission with seasonal peaks for both clinical malaria and hospital admissions with malaria and consistent results were observed when more than one outcome or more than one calendar year was available from the same site.

As their conclusion, shows that the proposed definition discriminated well between studies with 'marked seasonality' and those with less seasonality and recommends further work is needed to explore the applicability of this definition on a wide-scale, using routine health information system data where possible, to aid appropriate targeting of interventions.

CHAPTER THREE
TECHNIQUES FOR MINING MALARIA DATA

Hand et al. [**21**] noted each data mining algorithm can be decomposed into four components such as model or pattern structure, interestingness measure (score function), search method and data management strategy.

Witten and Frank [**14**] mentioned there exist many classification algorithms inside the WEKA system. WEKA (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms and data processing tools implemented in Java in 1993. It's developed as support for the whole process of experimental data mining such as preparation of input data, statistical evaluation of learning schemes, visualization of input data and the result of learning and used for education, research and applications. It's main features are 49 data preprocessing tools, 76 classification/regression/MLP algorithms, 8 clustering algorithms, 15 attribute/subset evaluators + 10 search algorithms for feature selection, 3 algorithms for finding association rules, 3 graphical user interfaces such as explorer (exploratory data analysis), experimenter (experimental environment) and knowledge flow (new process model interface) [**52**].

The researcher interested to explain the model selected and appropriate for this research. As mentioned in chapter one in the methodology sections, the question that this research is going to answer is a classification and association rule problems. For this reason, it is important to justify the model building and experiments to be carried out in the knowledge discovery process, which also involve data mining tool selection and algorithms used for modeling.

As a result, in this chapter, it is essential to discuss briefly on decision tree and rule induction classification as well as neural networks and how to increase the classification performance of the model as well as association rule mining and how it finds out the frequent itemset. Knowledge generated by data mining techniques can be represented in many different ways of which classification and association rule are commonly motioned for predictive and descriptive data mining modeling respectively.

3.1. Classification Model Techniques

For the classification purpose, Decision tree (J48) to generate rules sets and Rule Induction Method (JRip) well as multiplayer perceptron implemented. Witten and Frank [**14**] express classification algorithm as rule induction and decision-tree algorithms in WEKA systems.

Rule induction algorithms generate a model as a set of rules. The rules are in the standard form of IF-THEN rules. Meanwhile, decision-tree algorithms generate a model by constructing a decision tree where each internal node is a feature or attribute. The leaf nodes are class outputs [**26**]. Where us multilayer perceptron [**14, 42**] builds "black box" models.

3.1.1. J48 Decision Tree Algorithm

Decision tree is a popular utility that involves decision based classification and adaptive learning over a training set [**50**].Whitten and Frank [**14**] also stated J48 algorithm of decision tree technique is one of classification and prediction algorithms which support both numeric and nominal predicators and nominal class attribute values.

The J48 algorithm [**26, 49 and 50**] is the WEKA implementation of the C4.5 top-down decision tree learner proposed by Quinlan in 1993. The algorithm uses the greedy technique and is a variant of ID3, which determines at each step the most predictive attribute, and splits a node based on this attribute. It deals with numeric attributes by determining where thresholds for decision splits should be placed.

Decision tree algorithm [**14**] take inputs, data partition, D, which is a set of training tuples and their associated class labels, attribute list, the set of candidate attributes and attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of splitting attribute and, possibly, either a split point or splitting subset. The output of the algorithm will be a decision tree. Hereunder, we try to discuss basic algorithm for inducing a decision tree from a training tuples [**14**].

1. *Create a node N;*
2. *if tuples in D are all of the same class, c then*
3. *return N as a leaf node labeled with the class C_i*
4. *if attribute_list is empty then*
5. *return N as a leaf node labeled with the majority class in D;// Majority voting*
6. *apply **Attribute_selection_method (D, attribute_list)** to find the "best" splitting_criterion;*
7. *label node N with splitting_criterion*
8. *if splitting-attribute is discrete-valued and multiway splits allowed then // not restricted to binary trees*
9. *attribute_list <-attribute_list-splitting-attribute; // remove splitting-attribute*
10. *for each outcome j of splitting_criterion partition the tuples & grow subtrees*
11. *let D_j be the set of data tuples in D satisfying outcome j; // a partition*
12. *if D_j is empty then*
13. *attach a leaf labeled with the majority class in D to nod N;*
14. *else attach the node returned by Generate_decision_tree (Dj, attribute_list) to node N;*
15. *endfor*
16. *return N;*

Attribute selection method specifies a heuristic procedure for selecting the attribute that best discriminates the given tuples according to class. The process of decision tree generation by repeatedly splitting on attributes is equivalent to partitioning the initial training set into smaller training sets repeatedly, until the entropy of each of these subsets is zero (i.e. each one has instances drawn from only a single class).

At any stage of this process, splitting on any attribute has the property that the average entropy of the resulting subsets will be less than (or occasionally equal to) that of the previous training set [**10, 14, 21**]. The 'entropy method' of attribute selection is to choose to split on the attribute that gives the greatest reduction in (average) entropy, i.e. the one that maximizes the value of information gain. The splitting criterion tells us which attribute to test at node N by determining the best way to separate or partition the tuples in D into individual classes and indicates the splitting attribute and may also indicate either a split point or a splitting subset.

This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found. The expected information needed to classify a tuple in D is given by:

$$Info\ (D) = -\sum_{i=1}^{k} p_i \log_2 (p_i)$$

Where pi is the probability that an arbitrary tuple in D; belongs to class C_i and is estimated by $|C_i, D| / |D|$. A log function to the base 2 is used, because the information is encoded in bits. Info (D) is just the average amount of information needed to identify the class label of a tuple in D. At this point, the information we have is based solely on the proportions of tuples of each class. Info (D) is also known as the entropy of D. Suppose we were to partition the tuples in database D on some attribute A having V distinct values, $\{a_1, a_2, \ldots, a_v\}$ as observed from the training data. If A is discrete-valued, these values correspond directly to the V outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets, $\{D_1, D_2 \ldots D_V\}$; where D_j contains those tuples in D that have outcome a_j of A. These partitions would correspond to the branches grown from node N. Ideally; we would like this partitioning to produce an exact classification of the tuples. That is, we would like for each partition to be pure. However, it is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class). The amount of information we would still need (after the partitioning) in order to arrive at an exact classification is measured by:

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} * Info(D_j)$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the j^{th} partition. $Info_A$ (D) is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information (still) required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

$$Gain_{(A)} = Info_{(D)} - Info_A(D)$$

Gain (A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, (Gain (A)), is chosen as the splitting attribute at node N. This is equivalent to saying that we want to partition on the attribute A that would do the "best classification," so that the amount of information still required to finish classifying the tuples is minimal (i.e., minimum Info A(D)).

In addition, in decision tree, the pruned tree has a hierarchy in that the most significant variable that used to discriminate the records is located at the top. It optimizes computational efficiency as well as classification accuracy. The process of pruning (post-pruning) traditionally begins from the bottom of the tree (at the child leaves), and propagates upwards. J48 algorithm recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible [**14, 49, 50**].

The overlying principle of pruning is to compare the amount of error that a decision tree would suffer before and after each possible prune, and to then decide accordingly to maximally avoid error. The metric used to describe possible error, denoted error estimate (E), is calculated with the

$$E= (e+1)/ (N + m)$$

where 'E' is Error estimate, 'e' is misclassified examples at the given node, 'N' is examples that reach the given node, and 'm' is all training examples [**21, 49 , 50**].

Applying pruning methods to a tree usually results in reducing the size of the tree to avoid unnecessary complexity (produces fewer, more easily and interpretable results) and to avoid over-fitting of the data set when classifying new data that means improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set) [**21, 49**].

In the Weka J48 classifier, lowering the confidence factor decreases the amount of post-pruning since the effectiveness labeled by the confidence factor. Post-pruning in the C4.5 algorithm is the process of evaluating the decision error (estimated percent misclassifications) at each decision junction and propagating this error up the tree [**49, 50**]. At each junction, the algorithm compares the weighted error of each child node versus and Misclassification error (if the child nodes were deleted and the decision nodes were assigned the class label of the majority class).

Weka J48 algorithm also has subtree replacement and subtree raising pruning [**52**]. Subtree replacement pruning in each node in a decision tree may be replaced with a leaf basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. Where us, subtree raising in which a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Subtree raising often has a negligible effect

on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that subtree raising can be somewhat computationally complex [**52**].

3.1.2. JRIP Rule Induction Algorithm

The second classification technique used in this research is rule induction is JRip. JRip implements a prepositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by Cohen as an optimized version of IREP. Ripper builds a rule set by repeatedly adding rules to an empty rule set until all positive examples are covered. Rules are formed by greedily adding conditions to the antecedent of a rule (starting with empty antecedent) until no negative examples are covered. After a rule set is constructed, an optimization post pass massages the rule set so as to reduce its size and improve its fit to the training data [**26**].

We prefer JRip over other rule induction algorithms; As Daud and Corne [**26**] survey on decision tree and rule induction existing algorithms indicates JRip result are better readability and accuracy among the entire existing algorithm. The most essential issue they noted are JRip has both the ability and potential to produce accurate but readable rules [**26**] i.e. Rules which are generated using JRip algorithm are more clear and understandable. The detail of the algorithm listed as follows. The algorithm Initialize RS = { }, and for each class from the less prevalent one to the more frequent one,

1. *Repeat 2 and 3 until the description length (DL) of the rule set greater than the smallest DL met so far for example: when the error rate >= 50%.*
2. *Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate).*
3. *Incrementally prune each rule and allow the pruning of any final sequences of the antecedents*
4. *After generating the initial rule set {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 2 and 3.*
5. *Delete the rules from the rule set that would increase the DL of the whole rule set if it were in it. And add resultant rule set to RS.*
6. *end of psuedocode*

To elaborate JRip rule induction generally follows three stages. The first stage, building stage performs grow and prune phase. Grow Phase, grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: **p (log (p/t)-log (P/T))**. Prune Phase, incrementally prune each rule and allow the pruning of any final sequences of the antecedents. The pruning metric is (p-n)/ (p + n) -- but it's actually 2p/ (p + n) -1, so in this implementation we simply use p/ (p + n) (actually (p+1)/ (p+n+2), thus if p + n is 0, it's 0.5). The second stage, optimization Stage, after generating the initial rule set {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/ (P+N).Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the rule set. After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again. The last stage, delete the rules from the rule set that would increase the DL of the whole rule set if it were in it. And add resultant rule set to RS.

3.1.3. Multilayer Perceptron

Within the framework of neural networks, the multilayer perceptron [**14**, **42**] is one of the most widely used problem-solving architectures in a great variety of areas. Its known to its proficiency as an universal approximator of non-linear relationships between data input and output. In addition, it is easy to use and apply.

Multilayer Perceptron is an advance on simple Perceptron and arose in response to some limitations found in the simple version of the architecture. In 1986, Rumelhart et al [**42**] formalized a method through which a neuronal network could learn the existing association between the input patterns and the corresponding outputs, utilizing more levels of neurons than Rosenblatt used to develop the Perceptron. This method, known as backpropagation (backward error propagation), is an extension to networks with intermediate layers (multilayer networks) and non-linear activation functions of the delta rule proposed by Widrow and Hoff to account for the error produced by exits from the network.

Very briefly, the workings of the backpropagation network (the algorithmic discussion) consists in learning from a set of input-output pairs by means of the following process

1. *Frst, an input pattern is applied as a stimulus for the first layer of neurons of the network*
    - *It continues propagating through all the adjacent layers until generating an output*
    - *the results obtained in the output neurons are compared with the desired output*
    - *an error value is calculated for each output neuron.*
2. *Next, these errors are transmitted backwards,*
    - *starting from the exit layer, toward all the neurons of the intermediate layer that contribute directly to the output*
    - *receiving the percentage of error that corresponds to the participation of the intermediate neuron in the original output.*
    - *This process continues, layer by layer, until all the neurons of the network have received an error that describes their relative contribution to the total error.*
3. *Based on the value of the error received,*
    1. *the weights of the connections between the neurons are readjusted.*
4. *Thus, the next time the same pattern occurs the output will be closer to the desired value and in this way the error decreases*.

In general, the sigmoid function gives extra information necessary for the network to implement the back-propagation training algorithm. Back-propagation works by finding the squared error (the Error function) of the entire network, and then calculating the error term for each of the output and hidden units by using the output from the previous neuron layer. The weights of the entire network are then adjusted with dependence on the error term and the given learning rate [**14**].

Like any other learning scheme, multilayer perceptrons trained with backpropagation may suffer from overfitting [**14**]. To alleviate this, it uses early stopping, works like reduced-error pruning in rule learners. The error on the holdout set is measured and the algorithm is terminated once the error begins to increase, because that indicates overfitting to the training data. Another method, called weight decay, adds to the error function a penalty term that consists of the squared sum of all weights in the network.

This attempts to limit the influence of irrelevant connections on the network's predictions by penalizing large weights that do not contribute a correspondingly large reduction in the error.

In successive cycles the parameters (learning rate and soomthing factor) of the network are adjusted until the error reaches a minimum.

A. Learning Rate

In neural networks training [**42, 52**] is carried out by comparing the neural network output and the actual value and by adjusting the weights depending on the error value. Learning rate determines how big a change must be made towards the correct value i.e. do we take a giant step towards the correct value (large learning rate) or small step (small learning rate). A very high learning rate is not preferred since there would be giant oscillation as the network makes large adjustments for one pattern and another large change for the next pattern. Bigus, suggested lower learning rate at the beginning [**42**].

B. Smoothing Factor (Momentum)

This is a parameter that goes hand in hand with learning rate. The momentum parameter causes the errors from previous training patterns to be averaged together over time and added to the current error [**42, 52**]. If the error on a single pattern forces a large change in the direction of the neural network weights, this effect can be mitigated by averaging the errors from the previous training patterns'.

3.2. Association Model

For the association purpose, a priori techniques selected and implemented.

3.2.1. Apriori Algorithm Techniques

The a priori algorithm follows a step of find the frequent item sets that have minimum support with a subset of a frequent item set must also be a frequent support i.e. if {AB} is a frequent itemset, both {A} and {B} frequent itemset as well as iteratively find frequent itemsets with cardinality from 1 to k (k-itemset). Finally it uses the frequent itemsets to generate the rules. The algorithm looks like

```
L1 = {large 1-itemsets};
for (k=2; L k-1#Ø, k++) do begin
        Ck = apriori - gen (Lk-1); // new candidates
        forall transactions t ϵ D do begin
            Ci = subset (ck, t) // candidate contain in t
            forall candidates c ϵ Ct do
              c.count ++;
        end
        Lk = (c ϵ ck c.count | minsup)
end
```

The key concepts are frequent item sets (the set of item which has minimum support, denoted by L_i for i^{th} – item set), a priori property (any subset of frequent item set must be frequent) and Join operation (to find L_k, a set of candidate k – item sets is generated by joining L_{k-1} with it self).

Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence.

The next top quality of Apriori algorithm to implement was it's achievement of good performance by reducing the size of candidate sets but in situations with very many frequent itemsets, large itemsets, or very low minimum support, it still suffers from the cost of generating a huge number of candidate sets and scanning the database repeatedly to check a large set of candidate itemsets as noted in the survey [**50**]. In fact, it is necessary to generate 2^{100} candidate itemsets to obtain frequent itemsets of size 100.

An additional justification for Apriori implementation will be the number of database scans for the candidate generation algorithm (Apriori) increases with the dimension of the candidate itemsets as well as the performance decrease with the support factor and the FP-growth algorithm needs at most two scans of the database and the performance of the FP-growth algorithm is not influenced by the support factor as noted by Gyorödi et al [**53**] in their comparative study of association mining algorithms. And they conclude, in other cases the algorithms without candidate generation Dynamic FP-growth and FP-growth behave much better.

3.3. SMOTE Algorithms Technique

Smote Synthetic Minority Over-sampling Technique (Smote) [**52**] is an over-sampling method. Its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. Thus, the over

fitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space. This situation can occur when interpolating minority class examples can expand the minority class clusters, introducing artificial minority class examples too deeply in the majority class space [**55, 56**]. The detail of the algorithm looks like

Algorithm *SM OTE (T, N, k)*
Input: *Number of minority class samples T;AmountofSMOTEN%; Number of nearest neighbors k*
O utput: *(N/100)* T synthetic minority class samples*

1. *(* If N i s le ss than 100%, randomi z e t he mi nori ty c l ass s ample s as only a random pe rce n t of t he m wi l l be S M O TE d. *)*
2. ***if*** *N<100*
3. ***then*** *Randomize the T minority class samples*
4. *T =(N/100)* T*
5. *N =100*
6. ***endif***
7. *N =(int)(N/100)(* The amou n t of S M O TE i s assu med t o be i n i nteg ral mu l ti ple s of 100. *)*
8. *k = Number of nearest neighbors*
9. *numattrs= Number of attributes*
10. *Sample[ ][ ]: array for original minority class samples*
11. *newindex: keeps a count of number of synthetic samples generated, initialized to 0*
12. *Synthetic[][]:arrayforsyntheticsamples (* Compute k neare s t n e i g hbors f or eac h mi n ori t y c lass sample only. *)*
13. ***for*** *i← 1* ***to*** *T*
14. *Compute k nearest neighbors for i, and save the indices in the nnarray*
15. *Populate(N, i, nnarray)*
16. ***endfor***

Populate (N, i, nnarray)(Fu nc ti on to g e n e rate t he synthe ti c s ample s . *)*

17. ***w hile*** *N =0*
18. *Choose a random number between 1 and k,callitnn. This step chooses one of the k nearest neighbors of i.*
19. ***for*** *attr← 1* ***to*** *numattrs*
20. *Compute: dif =Sample[nnarray[nn]][attr]- Sample[i][attr]*
21. *Compute: gap= random number between 0 and 1*
22. *Synthetic[newindex][attr]=Sample[i][attr]+gap* dif*
23. ***endfor***
24. *newindex++*
25. *N =N- 1*
26. ***endw hile***
27. ***return*** *(* E n d of P opu late . *)*

End of Pseudo-Code.

The alogrthimic process is illustrated in Figure 3.3, where x_i is the selected point, x_{i1} to x_{i4} are some selected nearest neighbours and r_1 to r_4 the synthetic data points created by the randomized interpolation. The implementation employed in this work

uses the Euclidean distance, and balances both classes to the 50% distribution. Figure 3.1 depicts the illustration of this in detail.

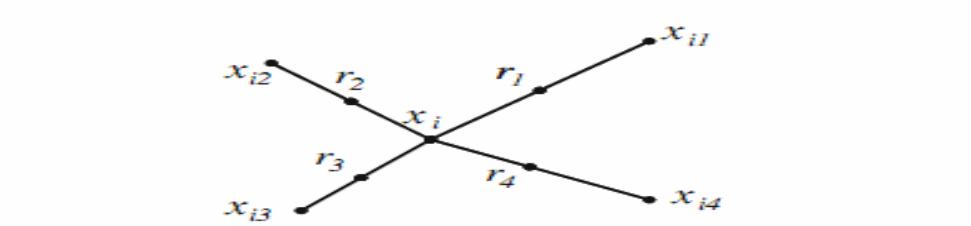


Figure 3.1. An illustration on how to create the synthetic data points using SMOTE algorithm

Figure 3.1 shows example of SMOTE application, consider a sample (6, 4) and let (4, 3) be its nearest neighbour. (6, 4) is the sample for which k- nearest neighbours are being identified and (4, 3) is one of its k-nearest neighbours. Table 3.1 explains the detail of Figure 3.3 SMOTE application with example.

Table 3.1 SOMTE Application Example

Let the values		Difference b/n feature sample	Result	New Sample Generated (f1',f2')
F1_1	6	F2_1 – F1_1	-2	(6, 4) * rand (0, 1) * (-2, -1)
F2_1	4			
F1_2	4	F2_2 – F1_2	-1	
F2_2	3			

Remark: Rand (0, 1) generates a random number between 0 and 1.

Synthetic samples are generated [**56, 57**] as shown in table 3.1: First, take the difference between the feature vector (sample) under consideration and its nearest neighbour. Then, multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general.

All in all, In SMOTE 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 neighbours [**56**]. Depending upon the amount of over sampling required, neighbours from the k nearest neighbours are randomly chosen.

3.4. Validation Techniques (Test Options)

For the experimental setup, the original malaria datasets are converted to ARFF (Attribute Relation File Format) as this is the input file format for the WEKA system.

Next, all the identified algorithms are tested to each malaria dataset with the option of using 10-fold cross-validation (the classifier evaluated using the number of folds that are entered in the folds text field).

10-cross validation used default parameters in the experiment and a standard way of predicting the error rate [**52**]. The k test instances have to be drawn off the training set. The bigger this set is, the more realistic is the estimate of the true error, but the less data is left over for use when training. A static division of the entire set of instances into training and test set may not be representative. This is why a so called cross-validation approach is applied. In cross-validation, the data is partitioned into a fixed number f of disjoint folds which are about the same size. A classifier is built f times; using f - 1 fold for training and one fold for testing.

At the end, every instance has been used once for testing. The test sets are independent, but the training sets which overlap are not independent of each other. The results of the evaluation are averaged over all f runs. A typical value for f is ten. Often, a process called stratification is used in conjunction with cross-validation. Stratification ensures that in each fold the original class distribution is maintained. This improves the stability of the evaluation results.

Also, 10-fold cross-validation noted that for this project data testing, we used is standard default setting inside WEKA system (version 3.7.3) without any modification [**52**].

In this test option the accuracy estimate is the overall number of correct classifications from the k iteration divided by the total number of samples, which is k. After deciding the values of the parameters the algorithm was run to start building the model.

By doing so the partition and the experiment could be more reliable. To illustrate the model, we have tested the system using a constructed dataset from malaria database, metrological and altitude dataset using WEKA 3.7.3 once the data set cleaned and prepared.

3.6. Evaluation Techniques

The basic measure is accuracy, which computes, the percent of correctly classified instances in the test set. Accuracy of a test compares how close a new test value is to a value predicted by if…then rules [**47**]. To classify a test example, the rule that matches it best determines the example's class membership. An accuracy test is defined as: Accuracy= (True Positive rate/Total number of test samples)*100%.

When the confusion matrix has only two outcomes (positive and negative) of a test are possible, three evaluation criteria can be used for measuring the effectiveness of the generated rules [**47**]. There are four possibilities, as shown in Table 3.2.

*Table 3.2 Possible outcome of the test set [**47**]*

	Test result Positive	Test result negative
Hypothesis positive	TP	FN
Hypothesis negative	FP	TN

Where TP, or true positive, indicates the number of correct positive predictions (classifications); TN or true negative is the number of correct negative predictions; FP or false positive is the number of incorrect positive predictions; and FN or false negative is the number of incorrect negative predictions [47]. The four measures are:-

1. Sensitivity = (TP/Hypothesis Positive)*100%= (TP/TP+FN)*100% i.e. the ability of a test to be positive when the condition is actually present or how many of the positive test examples are recognized.
2. Specificity= (TN/Hypothesis Negative)*100%= (TN/FP+TN)*100% i.e. the ability of a test to be negative when the condition is actually not present or how many of the negative test examples are excluded.
3. Predictive Accuracy = (TP+TN/ Total)*100%= (TP+TN/TP+TN+FP+FN)*100% i.e. a high level of confidence can be placed only for results that give high values for all three measures.
4. Precision= TP/ (TP+FP)*100%, how many of the test correctly classified from the total test and Recall= TP/(TP+FN)*100%, how many of the actual correct value classifies correctly.

CHAPTER FOUR

EPIDEMIOLOGY OF MALARIA AND DATA PRE-PROCESSING

Incidence and prevalence are two common epidemiological measures of endemicity [**44, 45**]. Incidence rate is the number of new cases of malaria per unit of population over a specific period of time (new occurrences of the disease over a period of time/ population at risk of the disease over that period of time) where as prevalence is the number of malaria at a point in time per unit of population. For example, assume you test a random sample of 1,000 people in one kebele for the presence of malaria parasites in their blood over a period of two weeks. If you find 500 of them to be positive for malaria, this means that the prevalence of parasitaemia in this population during the period of the survey was 50% (500/1000).

4.1 Determinant of Transmission

Malaria consortium [**46**] describes factors that determine the transmissions of malaria are the following. First, environmental factors which includes rainfall, temperature and humidity and altitude. Secondly, host factors such as age and immunity. Thirdly, genetic factors which include the sickle cell trait and Duffy blood group. Fourthly, social/ behavioral factors consisting of sleeping outdoors, working at night, house structure, traditional beliefs with treatment sought from traditional healers and leading to delays in care seeking, ability to afford protective measures or treatment, agricultural or industrial activities, wars and displacement. Finally, vectors and parasite factors such as plasmodium species, strain of species: virulence, resistance, vector behavior. Here below the researcher try to review detail description of each factors.

1. **Rainfall:** Important to the mosquito lifecycle and closely related with malaria seasonality.

*Table 4.1 Presents Condition of Rainfall for Malaria Transmission [**46**].*

Condition	**Description and Implication**
No rainfall	- Eggs will dry up and no breeding sites
Heavy rain fall and flooded rivers eggs	-Eggs and larvae may be washed away
Consistent rainfall	-Result puddles and breeding sites for mosquitoes

2. **Temperature and Humidity**: Temperature influences both the mosquito vector and malaria parasite and Humidity influence mosquito vector. Detail is shown in table 4.2 [**46**].

*Table 4.2 Presents Condition of Temperature and Humidity for Malaria Transmission [**46**].*

Condition	Description and Implication
High or hot Temperature (around 40^{o}c) and Humidity	-Life cycle of mosquito shorten -Shorter adult life span -Less malaria transmission -Change people behavior
25-32^{o}C	-Ideal temperature range -Mosquito develop with in shortest time
>35^{o}C	-Development will be impaired
< 15^{o}C for P. vivax and 20^{o}C for P. facliparum	-Sporogony will not be completed -Malaria transmission will not occur
Temperature drops	-Lengthen development time from egg & adult -Shortens adult life spans -Malaria transmission declines
Relative Humidity exceeding 60^{o}c	-Increasing the longevity of adult vectors -Anophelines usually survive best

3. **Altitude**: temperature decreases as altitude increases that means transmission is lower at high altitudes.

*Table 4.3 Depicts Altitude for Malaria Transmission in Ethiopia [**6**].*

Condition	Description and Implication
>2500m	Malaria free highlands
>=2000 and <=2500m	Highlands affected by occasional epidemics
<1500m with rainfall <500mm	Arid lowlands affected by occasional epidemics malarious near water
>=1750 and <2000m	Highland fringes with low transmission epidemic prone
>=1500 and <1750m	Highland fringes with high transmission epidemic prone
<1500m with rainfall >=500 and <=1000mm	Malarious lowlands with seasonal transmission
<1500m with rainfall >1000mm	Malarious lowlands with intense transmit ssion

4. **Age and Immunity**:

*Table 4.4 Presents Conditions Age and Immunity for Malaria Transmission [**46**].*

Condition	**Description and Implication**
High transmission area	-Children can be carrying around malaria parasites in their blood with becoming ill - Low level of asymptomatic parasitaemia in adult and old children b/c semi-immunity develops at several stages
Low transmission area and people don't develop immunity	-Fewer people carry malaria parasites in their blood i.e. "reservoir" of malaria gametocytes is higher where immunity exists
Low transmission area	-A person has less chance that will be bitten -Don't build up any immunity or develop immunity more slowly - Not only Childs and Pregnant Vulnerable but also age under 20 years

5. **Genetic Factors:** Every person is either positive or negative Duffy blood group based on the outer coats of the red blood cells.

*Table 4.5 Genetic Factors for Malaria Transmission [**46**].*

Condition	**Description and Implication**
Negative for Duffy Blood Group	-Resistant to invasion by P. vivax malaria, providing protection from vivax illness. -Most African are Duffy negative

6. **Social and Behavioural Factors:** These factors influence malaria transmission through increasing man vector contact and treatment seeking behaviour.

*Table 4.6 Social and Behavioral Factors for Malaria Transmission [**46**].*

Condition	**Description and Implication**
Sleeping outdoors	-Increase contact between the human and the mosquito, Lack of normal protection, Happened in hotter months which may coincide with the malaria transmission and Ethiopian bites both in indoors and outdoors
Working at night	-May be they are not sleeping under a net and -More likely to be bitten e.g. awake at night to look animal may be more at risk
House structure	-Create barrier and provide entrance for the mosquito
Traditional belief	-Treatment sought from traditional healers and leading to delays in care seeking
Income	-Ability to afford protective measures and treatment
Activities	-Agricultural and Industrial activities e.g. irrigation
Wars and Displacement	-Increase malaria transmission e.g. poor quality of house, health systems may be broken
Migration	-Internal Migration e.g. from high populated highland areas to fertile lowland areas

7. Vector and Parasite Factors:

*Table 4.7 Give a picture of Vector and Parasite Factors for Malaria Transmission [**46**].*

Condition	**Description and Implication**
Plasmodium Species	-Different species causes different level of illness -Different threshold of temperature for development
Strain of Species	-Virulence (occur with in the same parasite species) -Drug Resistance (strain specific with some parasite strains have not for others)
Vector Behavior	-Exophagy or Endophagy, mosquito prefers to feed outdoors or indoors -Exophily or Endophily, mosquito prefers to feed on animals or humans -Frequency or Blood feeding- some species take a blood meal every 2 days or 3 days
Breeding site preferences	-Pattern of Transmission

4.2 Strategies Used

WHO Regional Committee for Africa [**47**] proposed epidemiological surveillance system for various levels of the health system to prop up their system. This system fill the gap of the health system on detecting epidemics, spread of diseases, human suffering and loss of lives. The committee noted the above causes are due to the weakness of data collection, analysis and use of information for action at all levels and lack of resources and awareness are critical indicator of weak surveillance system. Thus, to prevent and combat epidemics of communicable diseases the committee proposed IDS system.

4.2.1 Integrated Disease Surveillance System (IDS)

WHO Regional Committee for Africa [**47**] established IDS form to avoid the problem of duplication of efforts and resources. IDS facilitate to improve failure of health workers to report first case of epidemic prone diseases as per standard, inadequate data collection, analysis, utilization and dissemination of surveillance data at district level. Those disease didn't get notice by most other system get attention through IDS this include the leading causes of childhood deaths in Africa e.g. pneumonia and diarrhoeal diseases, surveillance of malaria also deficient. IDS also provide centre of attention to solve inadequate attention given for evaluation of programme and involvement of laboratories using surveillance data and inadequate supervisory support as well as completeness and timelines of reporting.

The main basis of integrated disease surveillance system is data collection for action i.e. only the data necessary for action is collected and processed and achieved with the guiding principles of usefulness, simplicity and flexibility of the system, orientation to a specific action and integration. The main source of surveillance data can be routine report, epidemic report, case based reports and sentinel site reports (e.g. malaria drug resistance) from country, inter country, region and district level [**47**].

WHO also noted, the processed information has been disseminated to the responsible body to take action or manage intervention programme. And also list out critical players like programme managers, community leaders and staff of other sector and nongovernmental agencies that are taking action as well as provide feedback for the generated report and encourage the transmit of information to the next higher level[**47**].

All in all the main strategy and goal is finding solution for the problem that identified in analysis and interpretation as well as take action accordingly.

4.3 Data Collection Process and Attribute Description

IDS system contains the principal causes of mortality, disability and morbidity in Africa. It enhanced the use of standard definition for notification and case base reporting on selected diseases.

The list of diseases in IDS form group into epidemic prone diseases, disease targeted for eradication, diseases targeted for elimination and other diseases of public health importance [**47**]. Among those groups the researcher is interested to discuss on the group of public health importance. These disease include Diarrhoea (< 5 year old children), pneumonia (< 5 year old children), HIV/AIDS, Malaria, Trypanosomiasis, TV and Onchocerciasis. Since one of the goal of public health is to protect and prevent the community from malaria. And preventing malaria give a support in reduction of child mortality (MDG 4) and improvement of maternal health (MDG 5). The output generate support this by delivering interesting rules through knowledge discovery techniques.

All in all, WHO noted type of disease, age, in pregnancy, outpatient cases, inpatient death and cases are some of the determinant factors for decision making, planning and solving the problem of a specific zone or region [**47**]. IDS form also take account of the above decisive features.

4.3.1 Process of Data Collection

As mentioned above, malaria is one of the diseases which is grouped under public health importance and the data collected using the surveillance of case based reports from the community and health facility. The dataset used for this research is collected from zonal health facility across all the regions of Ethiopia by using monthly routine report. The district health office compiles all reports from health facilities and forward to the national level. The national level compiles all reports and sends them to WHO country office and regional office.

4.3.2 Description of Malaria Attribute in IDS Form

The Integrated Disease Surveillance malaria case based monthly routine report form contains the following main cases besides the general attribute. These are:

1. Total malaria outpatient cases, inpatient cases and inpatient deaths both in less than and greater than five years cases
2. Inpatient cases and deaths less than and greater than five year with severe anemia cases
3. Malaria in pregnancy outpatient cases, inpatient cases and deaths and
4. Lab confirmed uncomplicated malaria outpatient cases of P. falciparum and P.vivax less than and greater than five years.

The detail statistical summary of the dataset collected from zonal health facility and constructed target dataset used for this research will discuss in section 4.4.

An essential requirement in data mining research project is data preparation and pre processing to reveal unexpected patterns from electronic health records [**11, 14**]. As is common in real world including health care, missing value, noisy, and inconsistent data are common for most databases. Even if validating electronic data for completeness, continuity, and accuracy is an ongoing process, in this research correcting missing or incorrect data done from the primary sources. The WHO malaria databases contains some minor errors during the data entry especially missing value and noisy value.

4.4. Initial Data Source Selection

The data sources for this research are WHO malaria data base, Ethiopian National Metrology Agency and Ethiopian Mapping Agency.

The data available at WHO data base are collected from each region of the country per month from zonal health facility service as shown in table 4.1 by using WHO IDS form. Some of these diseases are malaria, pneumonia, diarrhea, AIDS and other diseases.

Table 4.8 WHO Zonal Center for Diseases Data Collection

REGION NAME	WHO ZONAL CENTER FOR DISEASES COLLECTION	TOTAL CENTER
ADDIS ABABA	LIDETA, AKAKI KALITI, CHERKOS,GULELE, YEKA, BOLE,ADDIS KETEMA, KOLFE KERANIO, ARADA and NEFAS SILK LAFTO	10
AFAR	AFAR 1, AFAR 2, AFAR 3, AFAR 4 and AFAR 5	5
AMHARA	AWI, BHAIR DAR, EAST GOJAM, WEST GOJAM, NORTH GONDER, SOUTH GONDER, OROMIA, NORTH SHEWA, NOTH WELLO, SOUTH WELLO and WAGHUMRA	11
BENISHANGUL GUMUZ	PAWE, METEKEL, KEMASHI, TONGO and ASOSA	5
DIRE DAWA	DIRE DAWA	1
HARARI	HARERI	1
GAMBELLA	AGNUAK, GAMBELLA SOUTH WESTERN, NUER and MEJENGER	4
OROMIA	EAST WELLEGA, EAST HARERGHE, WEST WELLEGA, JIMMA ILLUBABOR, EAST SHEWA, NORTH SHOA, SOUTH WEST SHEWA WEST SHEWA, WEST HARERGHE, GUJI, ARSI, BALE and BORENA	14
SNNPR	GEDEO, GURAGHE, GAMO GOFA, KEFA, DERASHE, ALABA, KEMBATA/ TEMBARO, KNOSO, YEM, SILTI, HADIYA, SOUTH OMO, AMARO, BENCH MAJI, WOLAYTA, DAWRO, SIDAMA, SHEKA , KONTA, BURJI and BASKETO	21
SOMALI	JIJIGA, AFDER, GODE, WARDER , DEGEHABUR KORAHE ,LIBEN and FIK	8
TIGRAY	SOUTH WESTERN TIGRAY, NORTH WESTERN TIGRAY , CENTRAL TIGRAY , EAST TIGRAY , SOUTH TIGRAY and MEKELLE	6
Total		86

There are 5160 records per individual dataset (i.e. malaria In pregnancy cases, severe anemia cases, malaria type cases and total malaria cases with a total of 18781 records) in the malaria database of WHO, collected from 2004 to 2009 from 86 zones across Ethiopia (i.e. since WHO collects data per months there are 86*12 months =1032 data with in one year). All the records are taken on the above specified year, since data

mining research can be conducted in huge number of records. Table 4.9 shows attribute used when collecting disease information.

Table 4.9 Detail Attributes in WHO Malaria Database

Name	NO. of Attributes	Attribute list		
General	5	Health Facility Name, Zone, Region, Year, Month		
Total Malaria	6	<5 years	Outpatient cases	
			In patient cases	
			Inpatient deaths	
		>5 years	Outpatient cases	
			In patient cases	
			Inpatient deaths	
In patient Malaria with Severe Anemia	4	<5 years	In patient cases	
			In patient deaths	
		> 5 years	In patient cases	
			In patient deaths	
Malaria in Pregnancy	4	Outpatient cases		
		In patient cases		
		Inpatient deaths		
Uncomplicated Malaria LAB Confirmed	4	<5 years	P. FALCIPARUM	Outpatient cases
			P. VIVAX	
		>5 years	P. FALCIPARUM	Outpatient cases
			P. VIVAX	
Total	22	-		

4.5 Description and Statistical Summary of Initial Dataset

This sub topic mainly lists the detail description and statistical summary of each attribute. WHO malaria data base contain 22 attributes. These are 5 basic and 17 detail attributes.

4.5.1 General Attribute

The general attribute mainly provide basic information about the geographic location and period of coverage. The attributes are listed below:

- **Country**: The Country where the malaria information is collected. Such as Ethiopia. The data type is nominal.
- **Region:** Administrative regions from which malaria information is collected. Values are ADDIS ABABA, AFAR, AMHARA, BENISHANGUL GUMUZ, DIRE DAWA, GAMBELLA, HARARI, OROMIA, SNNPR, SOMALI and TIGRAY. The data type is nominal.
- **Zone and Health Facility Name**: Specific zone of the region and each health facility across the zone where the data is collected. Values are name of the zone under the specific region. For example:- In Afar region there are 5 zones such as After 1, Afar 2, Afar 3, Afar 4 and Afar 5. Each has a number of health facility names. The data type is nominal.
- **Year**: The specific year in which malaria information is colleted. For example: - values of the year from 2004 to 2009. The data type is numeric.
- **Month**: The specific month of the year in which malaria information collected. For example: -January --- 01, February --- 02. The data type is nominal.

4.5.2 Detail Attribute

This attributes mainly describe and numerate the detail information of malaria in each zone of the region across Ethiopia in different categories as per WHO standards. The main categories are by age (less than, equal or greater than 5 year), malaria type (P. VIVAX and P. FALCIPARUM), cases (inpatient and outpatient), inpatient cases (cases and deaths), severe anemia (inpatient malaria cases less than 5 years and greater than 5 years), Lab confirmed uncomplicated malaria less than 5 years and greater than 5 years (P. VIVAX outpatient cases and P. FALCIPARUM outpatient cases). Here below list detail description and statistical summary of each attribute in WHO malaria database in Ethiopia and categorized the description in to four main categories.

- **Yes**: indicates the number of cases occurred i.e. the value is start from 1 to any number of cases or deaths (1 to n, n can be any positive integer).
- **No**: indicates there is no occurrence of the cases i.e. the values is 0.
- **Missing Value**: those blank values in the data base.
- **Noisy values**: unwanted values entered in the data base.

I. Malaria <5 years P.VIVAX.

Malaria less than 5 years P. VIVAX stands for uncomplicated malaria less than 5 years lab confirmed with which malaria type of P.VIVAX and it is numeric type attribute. Summary is presented in table 4.10.

Table 4.10 Summary of malaria < 5 years P.VIVAX

Uncomplicated Malaria <5 years P.VIVAX Lab Confirmed				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	3431	1257	15	1
Percent	72.93%	26.72%	0.0003%	0.0002%

II. Malaria <5 years P.FALCIPARUM

Malaria less than 5 years P. FALCIPARUM stands for uncomplicated malaria less than 5 years lab confirmed with which malaria type of P. FALCIPARUM and it is numeric type attribute.

Table 4.11 Summary of malaria < 5 years P. FALCIPARUM

Uncomplicated Malaria <5 years P.FALCIPARUM Lab Confirmed				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	3651	1040	12	1
Percent	77.61%	22.11%	0.003%	0.0002%

III. Malaria >5 years P.VIVAX

Malaria greater than 5 years P. VIVAX stands for uncomplicated malaria greater than 5 years lab confirmed with which malaria type of P.VIVAX and it is numeric type attribute.

Table 4.12 Summary of malaria >5 years P. VIVAX

Uncomplicated Malaria >5 years P.VIVAX Lab Confirmed				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	4050	652	11	1
Percent	86.09%	13.86%	0.23%	0.0002%

IV. Malaria >5 years P.FALCIPARUM

Malaria greater than 5 years P. FALCIPARUM stands for uncomplicated malaria greater than 5 years lab confirmed with which malaria type of P. FALCIPARUM and it is numeric type attribute.

Table 4.13 Summary of malaria >5 years P. FALCIPARUM

Uncomplicated Malaria >5 years P.FALCIPARUM Lab Confirmed				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	4175	517	11	1
Percent	88.75%	10.99%	0.23%	0.0002%

V. Malaria in Pregnancy In-Patient Cases

Malaria in pregnancy inpatient cases stands for those pregnant women who can get care in the health center getting a chance of bed admission and it is numeric type attribute.

Table 4.14 Summary of malaria in pregnancy inpatient cases

Uncomplicated Malaria <5 years P.FALCIPARUM Lab Confirmed				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	999	3664	40	1
Percent	21.23%	77.89%	0.008%	0.0002%

VI. Malaria in Pregnancy In-Patient Deaths

Malaria in pregnancy Inpatient deaths stands for those pregnant women who can get care in the health center by getting a chance of bed admission and dies while getting inpatient care and it is numeric type attribute.

Table 4.15 Summary of malaria in pregnancy inpatient deaths

Uncomplicated Malaria <5 years P.FALCIPARUM Lab Confirmed				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	105	4555	43	1
Percent	2.23%	96.83%	0.91%	0.0002%

VII. Malaria in Pregnancy Out-Patient Cases

Malaria in pregnancy outpatient cases stands for those pregnant women who can get care in the health center without getting a chance of bed admission (as a visitor only) and it is numeric type attribute.

Table 4.16 Summary of malaria in pregnancy outpatient cases

Uncomplicated Malaria <5 years P.FALCIPARUM Lab Confirmed				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	3036	1649	18	1
Percent	64.54%	35.06%	0.38%	0.0002%

VIII. Inpatient Malaria with Sever Anemia <5 years Cases

Inpatient malaria with sever anemia stands for those patients who can get care in the health center by getting a chance of bed admission with sever anemia case whose age less than five years and it is numeric type attribute.

Table 4.17 Summary of Inpatient Malaria with Severe Anemia <5 years case

In-Patient Malaria with Sever Anemia <5 years Cases				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	1045	3618	40	1
Percent	22.21%	76.91%	0.008%	0.0002%

IX. In-Patient Malaria with Sever Anemia <5 years Deaths

Inpatient malaria with sever anemia stands for those patients who can get care in the health center by getting a chance of bed admission with sever anemia case whose age less than five years and dies while getting patient care and it is numeric type attribute.

Table 4.18 Summary of Inpatient Malaria with Severe Anemia <5 years deaths

In-Patient Malaria with Sever Anemia <5 years Deaths				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	214	4450	39	1
Percent	4.55%	94.60%	0.008%	0.0002%

X. In-Patient Malaria with Sever Anemia >5 years Cases

Inpatient malaria with sever anemia stands for those patients who can get care in the health center by getting a chance of bed admission with sever anemia case whose age greater than five years and it is numeric type attribute.

Table 4.19 Summary of Inpatient Malaria with Severe Anemia > 5 years Cases

In-Patient Malaria with Sever Anemia >5 years Cases				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	1337	3325	41	1
Percent	28.42%	70.68%	0.0087%	0.0002%

XI. In-Patient Malaria with Sever Anemia >5 years Deaths

Inpatient malaria with sever anemia stands for those patients who can get care in the health center by getting a chance of bed admission with sever anemia case whose age greater than five years and dies while getting patient care and it is numeric type attribute.

Table 4.20 Summary of Inpatient Malaria with Severe Anemia >5 years deaths

In-Patient Malaria with Sever Anemia >5 years Deaths				
Cases	Yes	No	Missing Value	Noisy Value
Frequency	323	4337	43	1
Percent	6.86%	92.19%	0.009%	0.0002%

4.6. Data Cleaning

Data cleaning routines attempt to fill missing values, smooth out noise while identifying outliers and correct inconsistencies in the data [**11**].

4.6.1 Missing Values

Missing value can handle by ignoring the tuple [**11**]. When the class label is missing, for example: when the task involves classification which is not effective unless the tuple contains several attributes with missing values. And also poor when the percentage of missing values per attribute varies considerably. The other option is filling manually the missing values. This is time consuming and not feasible for large data set with many missing values. The third method is using a global constant to fill the missing value by replacing all missing attributes values by the same constant like "unknown". It is simple and not foolproof. Finally, filling missing values using the attribute mean (average) for all samples belonging to the same class as the given tuple and using the most probable value to fill in the missing value (determined with regression, inference-based tools using a Bayesian formalism or decision tree induction) [**11**].

To handle missing value the researcher tries to use and test both *global constant* and *most probable value* method.

4.6.2 Outliers Values

Outliers may be detected by clustering (similar values are organized in to groups or clusters and values fall outside the set of clusters may be considered outliers)[**11**]. The data used for this research is free of outliers. To detect outlier values a common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5*IQR above the third quartile (Q3) or below the first quartile (Q1) as stated by Han and Kamber [**11**]. In other words it is to mean that the values outside the limits: Q3 + (1.5*IQR) and Q1-(1.5*IRQ) will be considered outlier values.

4.6.3 Noisy Values

Noise is a random error or variance in a measured variable [**11**]. The noise can be removed by smoothing the data through binning (smooth a sorted data value by consulting its "neighbourhood".

The noisy value entered in WHO database is "UncomplicatedMalPVivax" across the entire record with out any value (the only one as shown in Description and Statistical Summary of Attributes). It is typing error which is out of WHO IDS (Integrated Disease Surveillance) standard form. Therefore to handle this problem the researcher automatically deletes this record.

4.7. Target Dataset Construction

In the fields of epidemiology and public health, the distinction between primary and secondary data depends on the relationship between the person or research team who collected a data set and the person who is analyzing it [**48**]. This is an important concept because the same data set could be primary data in one analysis and secondary data in another. If the data set in question was collected by the researcher (or a team of which the researcher is a part) for the specific purpose or analysis under consideration, it is primary data. If it was collected by someone else for some other purpose, it is secondary data.

Boslaugh [**48**] also noted advantage of working with secondary data. Since it is economical, breadth of data available and data collection process is informed by expertise and professionalism that may not available to smaller research projects. However, secondary data have demerit of inherent in its nature (i.e. In any case, you can only work with the data that exist, not what you wish had been collected to answer your specific research question) and the analyst did not participate in the planning and execution of the data collection process, he or she does not know exactly how it was done.

The researcher tries to make sense the secondary data by extracting target dataset from the data base (i.e. WHO malaria database and National Metrological data). From the available secondary data sets the researcher seeks target dataset that will allow for analysis of the research to answer the research question. This method conforms more to standard beliefs about how research is done and produce quality research.

Here below the researcher tries to construct or create a list of target datasets from WHO malaria dataset that include information related to the research question and achieve the objective of the research.

A. Total Malaria Cases and Deaths

Summary of total malaria cases and deaths dataset presented in table 4.21.

Table 4.21 Total malaria cases and deaths

Initial Dataset Attribute	**Target Dataset (Attribute Constructed)**
Total malaria inpatient < 5 years deaths	- Age - Type of Malaria Visits - Total number of cases - Total number of deaths
Total malaria inpatient < 5 years cases	
Total malaria inpatient > 5 years deaths	
Total malaria inpatient > 5 years cases	
Total malaria outpatient <5 years cases	
Total malaria outpatient >5 years cases	

From table 4.22 a total of four attributes constructed. The value of the target data set include age (less than and greater than 5), Type of malaria visits (inpatient and outpatient), total number of cases (numeric) and total number of death (numeric).

***B.* Uncomplicated Malaria Lab Confirmed (by malaria type)**

Table 4.22 Malaria with severe anemia

Initial Dataset Attribute	**Target Dataset (Attribute Constructed)**
Uncomplicated malaria < 5 year PF cases	- Age -Type of Cases - Type of malaria visit - Type of malaria - Number of cases
Uncomplicated malaria < 5 year PV cases	
Uncomplicated malaria >5 year PF cases	
Uncomplicated malaria >5 year PV cases	

From table 4.22 a total of five attributes constructed. The value of the target data set are age (less than and greater than 5), Type of Cases (Uncomplicated Lab Confirmed), Type of malaria visits (inpatient and outpatient), Type of malaria (PV, PF) and number of cases (numeric).

C. Malaria In pregnancy

Table 4.23 depicts inpatient or outpatient cases of malaria in pregnancy cases and deaths.

Table 4.23 Malaria in pregnancy

Initial Dataset Attribute	Target Dataset (Attribute Constructed)
Malaria in pregnancy in patient cases	-Age -Type of Cases -Type of Malaria Visits -Number of Cases -Number of Deaths
Malaria in pregnancy in patient deaths	
Malaria In pregnancy out patient cases	

From table 4.23 a total of five attributes constructed. The value of the target data set are age (Greater than 5), Type of cases (In pregnancy), Type of malaria visits (inpatient and outpatient), number of cases and deaths (numeric).

D. Malaria with severe anemia

Summary of inpatient malaria with severe anemia less than or greater than five years are presented in table 4.24

Table 4.24 Malaria with severe anemia

Initial Dataset Attribute	Target Dataset (Attribute Constructed)
Inpatient malaria with severe anemia < 5 years cases	- Age - Type of cases - Type of malaria - Type of malaria visits - Number of cases - Number of deaths
Inpatient malaria with severe anemia < 5 years deaths	
Inpatient malaria with severe anemia > 5 years cases	
Inpatient malaria with severe anemia >5 years deaths	

From table 4.24 a total of six attributes constructed. The value of the target data set are age (less than and greater than 5), Type of cases (malaria with severe anemia), Type of malaria visits (inpatient), Type of malaria (unknown, number of cases) and deaths (numeric).

In conclusion from total malaria cases and deaths, uncomplicated malaria lab confirmed (by malaria type), malaria in pregnancy and malaria with severe anemia data set a total of 8 attributes were constructed to achieve the research objective. These are age, type of malaria, type of cases, type of malaria visits, number of cases, number of deaths, total number of cases and deaths. The research tries to discuss

integration of these records as well as integration of metrological data with the malaria dataset in the next session.

4.8 Data Integration

Data need to be transformed into appropriate data integration format from multiple databases, data cubes or flat files by considering schema integration and object matching that are appropriate for mining [**11**].

To achieve the research objective, the researcher tries to integrate from different sources as well as between flat files as listed below.

A. Integration Between Flat Files

Based on the entity constructed and identified in section 4.4, the researcher tries to integrate files to achieve the researcher objective. Besides the common attribute country, regions, zones, year and month used for integration as shown in table 4.18

Table 4.25 Malaria Severe Anemia Data Integration

No.	Name of Flat File	kind of Cases	No. of Records	Entity used for Integration	Total no. of record after Integration
1	Inpatient Malaria with Severe Anemia	<5 yrs	4703	Age, Type of malaria, Type of malaria visits, Type of cases, Number of cases and Number of deaths	9403
		>5 yrs	4703		
2	Malaria In pregnancy	Outpatient cases	4703		9403
		Inpatient Cases and Deaths	4703		
3	Uncomplicated malaria lab confirmed (Malaria by Type)	<5 yrs PF	4703		18, 805
		>5 yrs PV	4703		
		< 5 yrs PV	4703		
		>5 yrs PF	4703		
4	Malaria by Cases	Inpatient Malaria with Severe Anemia	9403		18, 805
		Malaria In pregnancy	9403		
5	Total Malaria Data	Malaria by case	18, 805		37, 609
		Malaria by Type	18, 805		

Table 4.25 demonstrates the integration of inpatient malaria with severe anemia less than 5 year and greater than 5 year inpatient cases and deaths. Secondly, illustrate the integration of in pregnancy outpatient cases and inpatient cases as well as deaths and uncomplicated malaria <5 year and >5 year with PV and PF integrated. Thirdly, the above two malaria cases (severe anemia and in pregnancy cases) also integrated based on the identified entity. Finally, the constructed final dataset integrated between malaria cases and malaria types based on the selected attribute.

B. Integration Between Data Sources

The researcher also integrates metrological data sources with the malaria database and Ethiopian mapping agency data sources by simply appending the entity to the malaria data base based on region, zonal, year and monthly information. These attribute are temperature, rainfall and altitude. The final record prepared for model implementation and experimentation is 37, 609.

4.9 Data Transformation and Reduction

Data aggregation and dimensionality reduction involve removing irrelevant attributes, data compression using encoding schema such as minimum length encoding and sampling) [**10, 11**]. The initial data set for this research project contain 31 attributes and the final target dataset constructed include only 15. The research projects focus on uncovering knowledge from the malaria data by using predicting the death and cases of malaria as well as exploring the pattern. So, some attributes that doesn't have relation to the specified objective were removed. Therefore, due to this data reduction process, a total of 15 attribute constructed that is 12 attributes are selected for data mining process and 2 attributes from National Metrological Agency and 1 attribute from National Mapping Agency.

During data transformation, the data are transferred or consolidated into forms appropriate for mining tool. All data transformation result is given in table 4.25 with their values.

Table 4.26 Summary of Transformed Dataset

S.NO.	Attribute	Values
1	Regions	ADDIS ABABA, AFAR, AMHARA, BENISHANGUL GUMUZ, DIRE DAWA, HARARI, GAMBELLA, OROMIA, SNNPR, SOMALI and TIGRAY
2	Zones	Name of the zone (string) of each region
3	Year	Y2004, Y2005, Y2006, Y2007, Y2008 and Y2009
4	Month	M1,M2,M3,M4,M5,M6,M7,M8,M9,M10,M11 and M12
5	Age	Under 5 and Greater 5
6	Type of Malaria Visits	In patient and Out patient
7	Type of Cases	In pregnancy , Severe Anemia and Un complicated Lab Confirmed
8	Type of Malaria	PV, PF and Not Known (i.e. in Case of Severe Anemia and In pregnancy malaria type not known or determined in the dataset)
9	Number of Cases	Numeric
10	Occurrence of Deaths	Not Probable (i.e. No death exist), Probable (Death exist) and Undetermined (In case of Outpatient visit in Pregnancy and uncomplicated lab confirmed death not known or listed in the dataset)
11	Total Number of Cases	Numeric
12	Total Number of Deaths	Numeric
13	Temperature	$0\text{-}5^0$c, $5\text{-}10^0$c, $11\text{-}15^0$c,$16\text{-}20^0$c, $21\text{-}25^0$c, $25\text{-}30^0$c, $31\text{-}35^0$c, $35\text{-}40^0$c and $>40^0$c transformed in to {T1,T2,T3,T4,T5,T6,T7,T8 and T9}
14	Rainfall	Numeric
15	Altitude	>3500m, 2500-3500m, 2000-2500m, 1500-2000m, 1000-1500m, 500-1000m, 0-500m and < zero transformed into {E1,E2, E3, E4, E5, E6, E7 and E8}

4.10 Summary of Initial and Target Dataset

The malaria dataset used for this study consists of the attribute Regions, Zones, Year (GC), Month Number (GC), Altitude, Temperature, Rainfall, Age, Type of Malaria Visits, Type of Cases, Type of Malaria, Number of Cases, Occurrence of Death, Total number of Cases and Deaths.

Table 4.27 presents the comparative summary of initial and target datasets. Showing reduction in the number of attributes and files size.

Table 4.27 Summary of initial and target Dataset

Summary	**Initial Dataset**	**Constructed Dataset/ Target Dataset (From three sources)**		
No. of Attributes	31	15		
File Format	.xls	.xls	.CSV	.arff
File Size	3.03 MB	8.54 MB	4.12 MB	4.00MB
Total no. of records	4,703 (No. records of one single independent flat file)	37, 609		

CHAPTER FIVE
PREDICTIVE MODEL EXPERIMENTATION

Weka J48 provides several other options that determine the specificity of the model as shown in Fig. 5.1 [**52**]. One of the opportunity to which dictate the lowest number of instances that can constitute a leaf is the minimum number of instances per leaf. The higher the number, the more general the tree is. Lowering the number will produce more specific trees, as the leaves become more granular.

The most basic parameter is the tree pruning option. If you decide to employ tree pruning, we may need to consider the options for pruning. It is important to know that depending on how the training and test data have been defined that the performance of unpruned tree may superficially appear better than a pruned one [**52**].

The other is the binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using an inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option effectively treats the data as a nominal value. Turning this encourages more generalized trees.

Laplace also determinant option for smoothing for predicted probabilities. Laplace smoothing is used to prevent probabilities from ever being calculated as zero. This is mainly to avoid possible complications that can arise from zero probabilities. The detail explanation of each parameter in weka J48 algorithm classifier listed in fig.5.1.

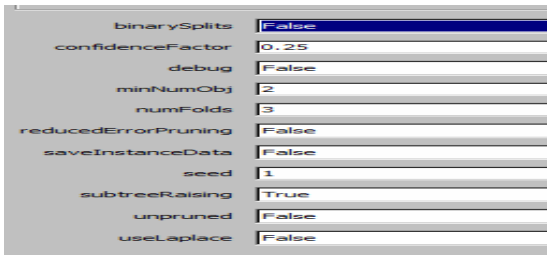


Figure5.1 Default J48 classifier Parameter option in weka

- BinarySplits: Whether to use binary splits on nominal attributes when building the trees.
- ConfidenceFactor: The confidence factor used for pruning (smaller values incur more pruning).
- Debug: If set to true, classifier may output additional info to the console.
- MinNumObj: The minimum number of instances per leaf

- NumFolds: Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.
- ReducedErrorPruning: Whether reduced-error pruning is used instead of C.4.5 pruning.
- SaveInstanceData: Whether to save the training data for visualization.
- Seed: The seed used for randomizing the data when reduced-error pruning is used.
- SubtreeRaising: Whether to consider the subtree raising operation when pruning.
- Unpruned: Whether pruning is performed.
- UseLaplace: Whether counts at leaves are smoothed based on Laplace.

A. Decision Tree Notation

The decision tree notation, as illustrated in a diagram of a decision in Figure 5.2, is called a decision tree. This diagram is read from left to right. The topmost node in a decision tree is called the root node. In Figure, this is a small square called a decision node. The branches emanating to the right from a decision node represent the set of decision alternatives that are available. One, and only one, of these alternatives can be selected. The small circles in the tree are called chance nodes. The number shown in parentheses on each branch of a chance node is the probability that the outcome shown on that branch will occur at the chance node. The bottom end of each path through the tree is called an endpoint, and each endpoint represents the final outcome of following a path from the root node of the decision tree to that endpoint.

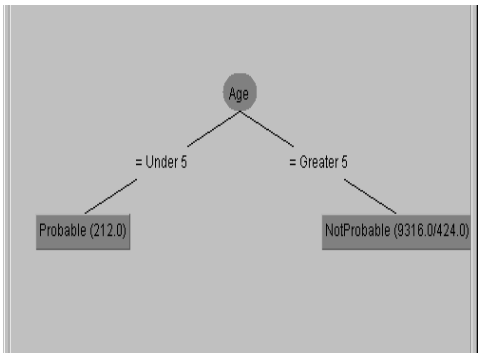


Figure 5.2 Decision Tree Diagram

JRip class capabilities may contain binary class, nominal class and missing class values. And the attributes can be missing values, unary attributes, numeric attributes, empty nominal attributes, date attributes nominal attributes and binary attributes [**53**].

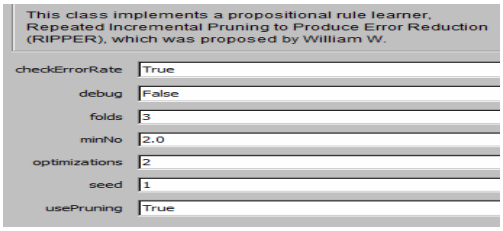


Figure 5.3 JRip Rule Induction Parameter Option

The JRip parameter indicates:-

- CheckErrorRate -- Whether check for error rate >= 1/2 is included in stopping criterion.

- Debug -- Whether debug information is output to the console.
- Folds -- Determines the amount of data used for pruning. One fold is used for pruning, the rest for growing the rules.
- minNo -- The minimum total weight of the instances in a rule.
- Optimizations -- The number of optimization runs.
- Seed -- The seed used for randomizing the data.
- UsePruning -- Whether pruning is performed.

Multiplayer Perceptron classifier that uses back propagation to classify instances can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the output nodes become unthresholded linear units) [**52**].

MLP [**52**] provides different parameter option such as GUI (Brings up a graphical user interface), auto build (Adds and connects up hidden layers in the network), debug (If set to true, classifier may output additional info to the console), decay (This will cause the learning rate to decrease and divide the starting learning rate by the epoch number, to determine what the current learning rate should be. This may help to stop the network from diverging from the target output, as well as improve general performance), hidden Layers (This defines the hidden layers of the neural network), learning Rate (The amount the weights are updated), momentum (Momentum applied to the weights during updating), nominalToBinaryFilter (This will pre process the instances with the filter and improve performance if there are nominal attributes in the data), normalizeAttributes (This could help improve performance of the network), normalizeNumericClass (This could help improve performance of the network, It normalizes the class to be between -1 and 1), randomSeed (Seed used to initialize the random number generator. Random numbers are used for setting the initial weights of the connections between nodes and also for shuffling the training data), reset (This will allow the network to reset with a lower learning rate. If the network diverges from the answer this will automatically reset the network with a lower learning rate and begin training again. This option is only available if the gui is not set. Note that if the network diverges but isn't allowed to reset it will fail the training process and return an error message), training time (The number of epochs to train through. If the validation set is non-zero then it can terminate the network early), validationSetSize

(The training will continue until it is observed that the error on the validation set has been consistently getting worse, or if the training time is reached) and validationThreshold (Used to terminate validation testing. The value here indicates how many times in a row the validation set error can get worse before training is terminated).

5.1. Visualization and Problem of Imbalanced Dataset

As shown in Figure 5.3 and 5.4, in the classification problem filed, as noted by Chawala et al [**55**] the scenario of dataset is imbalanced if the classification categories are not approximately equally represented. The researcher focuses on the occurrence of death class with the class value of undetermined class (majority), probable and non probable class (minority) as well as since uncomplicated lab confirmed on type of class case is the majority one.

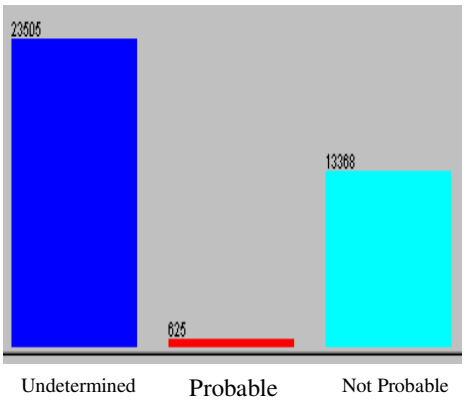


Figure 5.4 Occurrences of Deaths

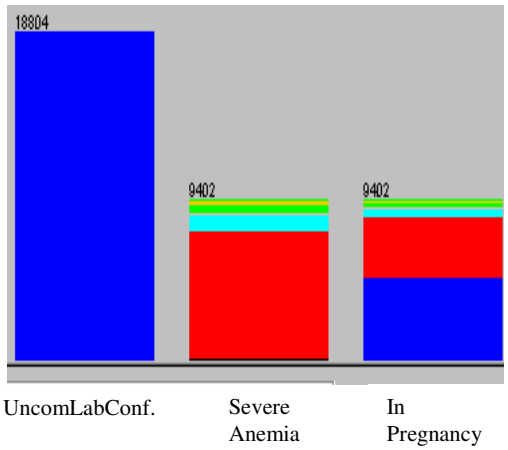


Figure 5.5 Types of Cases

Most of the learning algorithms aim to obtain a model with high prediction accuracy and a good generalization capacity [**57, 58**]. However, this inductive bias towards such a model supposes a serious challenge with the classification of imbalance data [**55**].

Foremost, if the search process is guided by the standard accuracy rate, it benefits the covering of the majority examples. Next, classification rules that predict the probable death class are often highly specialized and thus their coverage is very low; hence they are discarded in favor of more general rules, i.e., those that predict the undetermined class. Furthermore, it is not easy to distinguish between noise examples and minority class examples and they can be completely ignored by the classifier.

Imbalance learning problem grows since it is a recurring problem in many applications. For this reason, a large number of approaches have been previously proposed to deal with the class imbalance problem.

These approaches generally can be categorized into internal approaches that create new algorithms or modify existing ones to take the class imbalance problem into consideration and external approaches that pre process the data in order to diminish the effect cause by their class imbalance [**58**].

Furthermore, cost-sensitive learning solutions incorporating both the data and algorithmic level approaches assume higher misclassification costs with samples in the minority class and seek to minimize the high cost errors.

The great advantage of the external approaches is that they are more versatile, since their use is independent of the classifier selected. Furthermore, the researcher pre process all data sets before-hand in order to use them to train different classifiers. In this manner, the computation time needed to prepare the data is only used once.

However, the synthetic minority over sampling technique (SMOTE) [**55, 56**] is an important approach by over sampling the positive class or the minority class to handle imbalance of class as shown in figure 5.6 and 5.7.

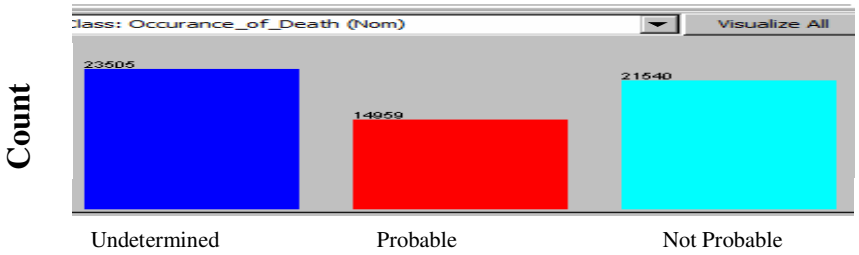


Figure 5.6 Occurrence of Death Class balanced using SMOTE

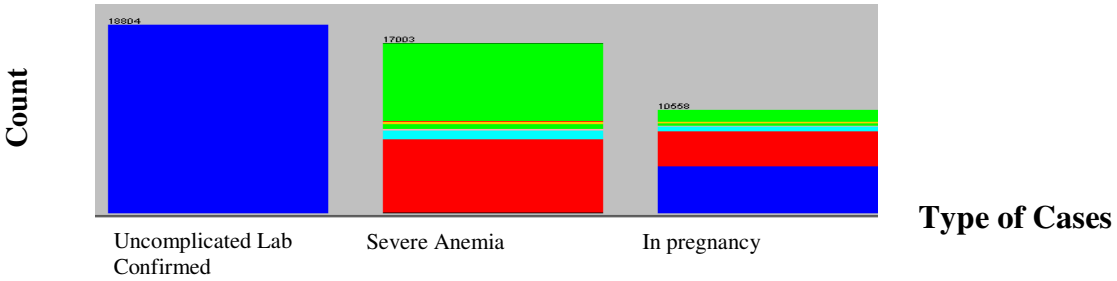


Figure 5.7 Type of Cases Class balanced using SMOTE

5.2. Experimental Scenario

To predict malaria occurrence of deaths and type of cases using J48, we used two scenarios (pruned and unpruned J48 decision tree algorithm techniques). WEKA 3.7 provides an option with use pruning value true or false. The default value is true.

In case of JRip, to predict occurrence of death and type of cases, we used two scenarios (pruned and unpruned JRip rule induction algorithm techniques). WEKA 3.7 provides an option with use pruning value true or false. The default value is true. In other words pruned JRip carry out when the value of use pruning is true and unpruned JRip execute when the value of use pruning false. A total of two experiments execute using JRip algorithm techniques and it generates 169 rules (45 rules using pruning and 124 rules with out pruning).

In conducting the experiment with multilayer perceptron, the amount the weights are updated was set to 0 .3 , the momentum applied to the weights during updating was set to 0 .2 , and the number of periods to train through was set to 500 epochs on each k-fold, any further training cycles may have induced over fitting.. Classifying the text using multilayer perceptron only succeed while using feature selection approach due to the computational complexity issue in the classifier .

5.3. Comparative Evaluation of Classification Rates

The performance of each algorithm has been evaluated under the following criteria in order to obtain an approximation to the performance capabilities of the algorithms for classification tasks in terms of accuracy (how well does each algorithm adapt to the training data or the mean percentage of correct classifications over each testing set in each of the folds), root mean square error (provides the mean error over each k folded testing set), kappa statistic (represents the mean agreement of the class predictions), where if a network predicts 0.1, and the actual classification is 0.1, the Kappa statistic will be 1.0. Non weighted KNN classification accuracy has been included in the comparison chart with k=1 as a separate performance measure and so on.

It is worth pointing out, that the initial objective of this study was not to discover which algorithm is superior in classification tasks, but to examine the advantages and downfalls of each algorithm under varying conditions to uncover knowledge in the malaria data. It is clear from brief inspection of the above results that each algorithm has varying performance over different datasets. It is also worth mentioning that

dataset pre-processing prior to training (where conversion is needed, i.e. from nominal form to numeric) could have an impact on the accuracy of the algorithm. Table 8.1 and 8.2 summarized the prediction performance of J48, JRip and MLP techniques.

Table5.1. Comparison of confusion matrix for prediction results

Techniques		TP Rate (%)	FP Rate (%)	Precision (%)	Recall (%)	F-Measure (%)	ROC Area (%)
J48	Occurrence of Deaths	95.9	2	95.9	95.9	96%	99
	Type of Cases	89.2	4.5	89.4	89.2	89.3%	98.4
JRip	Occurrence of Deaths	95.6	2.3	95.6	95.6	95.5%	97.9
	Type of Cases	80.6	13.1	85.6	80.6	79.8%	86.6
MLP	Occurrence of Death	97.4	1.1	97.4	97.4	97.4	99.2
	Type of Cases	87	6.4	86.9	87	86.6	95

To analyze how classification errors are distributed among classes as shown in table 5.1 we computed cumulative confusion matrices by observation of confusion matrices of each model. The J48 and MLP reveal that the trained on average generalized with lower error to the k-fold validation sets.

Table5.2. Comparison of training time and accuracy results

Techniques		Time (seconds)	Model Accuracy (%)	
			Correctly Class. Instances	Incorrectly Class. Instances
J48	Occurrence of Deaths	3.07	96.003	3.9997
	Type of Cases	1.61	89.2268	10.7732
JRip	Occurrence of Deaths	195.05	95.5736	4.4264
	Type of Cases	65.72	80.5629	19.4371
MLP	Occurrence of Deaths	108286.77	97.4185	2.5815
	Type of Cases	1440.46	86.9772	13.0228

All algorithms techniques performed well over the dataset, with J48 techniques gaining a slight classification accuracy advantage over the JRip rule induction and back propagation gaining a slight classification accuracy advantage over the J48 as

listed in table 5.2 In terms of training time, although not part of my assessment criteria, but it may be appropriate to say that for this particular class of classification problem, or at least this training set, decision tree induction may be a more suitable candidate.

By examining the confusion/contingency matrix it is possible to see that J48 and MLP demonstrate comparable classification accuracy on this dataset. The ROC (Receiver Operating Characteristics) area of J48 (99%) and MLP (99.2%) model is highest. The higher numbers indicates the model is the best accurate than the others. The ROC curve is a plot of how the classifier performs over the entire range of possible choices of cut-off values.

5.4. Summary and Analysis of the Result

5.4.1. Rules Generated using J48

A. J48 rule indicate occurence of deaths

J48 generate different interesting rules that predict occurrence of deaths. It generates excepted result. For instance in case of out patient malaria visits it's difficult to predict the death (i.e. if type of malaria visit is out patient it is difficult to determine occurrence of death) may be this is the limitation of the dataset. There is a high probability of occurrence of death when the age is below 5. However, J48 discover rules that indicate the probability occurrence of death in the following inpatient case scenarios.

Scenario I, In the case of inpatient malaria visits occurrence of death happened in the following scenarios i.e. if Type_of_Malaria_Visits = Inpatient and Number_of_Cases <= 12 and Number_of_Cases > 0 and zone

- Cherkos and Year (GC) and 2004 and Rainfall >5
- BAHIR DAR and Number of Case b/n 11 and 12, Year =2005, Number of cases > 12 and Year 2006
- W GOJJAM and MonthNumber (GC)= M11 and
  - ✓ Year (GC) = 2004 and Number of Cases <= 6 or >12
  - ✓ Year GC(2005) and Number of Cases <= 4
- BALE and Temperature= T5 and Number of Cases<= 3
- HORO GUDURU and Age = Greater 5 (not below 5)
- GURAGHE and Temperature and T4 and Age = Under 5 and Rainfall > 185 and Number_of_Cases > 3

- GURAGHE and Temperature =T5 and Type of Cases= SevereAnemia and age= greater 5
- KEFFA and Number of Cases > 1.495456
- AMARO and MonthNumber (GC) and M9 and M10
- DAWRO and Age =Under 5
- BASKETO and Number of Cases > 3.496872
- LIBEN and Number of Cases > 5.49636
- MEKELLE and Age and Under 5 and Year
    - ✓ Y2005 and No_of_Cases > 2
    - ✓ Y2006 and No_of_Cases > 3
    - ✓ Y2007

Scenario II, In case of Inpatient malaria visits when the number of case become larger and larger there is a probability of occurrence of death. The scenario looks like in specific region as follows. I.e. if Type_of_Malaria_Visits = Inpatient and Number_of_Cases > 12 and Regions

- AMHARA and Year (GC) = Y2004 and Number of Cases
    - <= 12.484261 and Rainfall > 15.001088
    - > 19.601178 and <= 21.30256 and Rainfall <= 15.9425
    - <= 21.350253 and Rainfall > 15.9425 and <= 16.240905
    - > 28.805297
    - <= 13.973823 and Zones = BAHIR DAR
    - <= 19 and Zones = W GOJJAM
    - > 19.164967 and Zones = N GONDER
- DIRE DAWA
- SNNPR and Number of Cases
    - > 17.957834 and Type_of_Cases = UncomplicatedLabConfirmed
    - > 17.957834 and Type_of_Cases = SevereAnemia and Age = Under 5 and Temprature and T5 and Number_of_Cases > 38.499341
    - > 17.957834 and Type_of_Cases = SevereAnemia and Age = Greater 5
- SOMALI
- TIGRAY and Age
    - Age = Greater 5 and Year

- ✓ Y2004: Probable

All other scenarios not listed here indicates occurrence of death not probable and unrelated. For example: if region is afar and zone says Addis Ababa (unrelated rules)

I. Analysis of the model (Confusion Matrix Analysis)

The experimental evaluation of the models/rules was taken place based on the performance/accuracy of models and confusion matrix, discussion with the domain expert and based on the soundness of the rule generated.

Calculating correctly classified instances and incorrectly classified instances is based on the confusion matrix as mentioned in methodology section. The detail of accuracy by class and confusion matrix presented as follow.

II. Summary

Table 5.3 Summary of J48 occurrence of death models

No.	Analysis		Pruned J48	Unpruned J48
1	Tree Size	No. of Leaves	498	2294
		Size of Tree	598	2563
2	Model Accuracy	Correctly Classified Instances	96.003%	95.8753
		Incorrectly Classified Instances	3.9997%	4.1247
3	Time		3.07 seconds	1.83
4	Kappa statistic		0.9388	0.9369
5	Mean absolute error		0.0379	0.0359
6	Root mean squarer error		0.1407	0.144
7	Relative absolute error		8.6615%	8.2189
8	Root relative squared error		30.0985	30.8138
9	Total Number of Instances		60004	

III. Pruned J48 Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	Undetermined
0.884	0.015	0.952	0.884	0.917	0.989	Probable
0.969	0.045	0.923	0.969	0.946	0.99	Not Probable
0.96	0.02	0.961	0.96	0.96	0.993	Weighted Avg.

The confusion matrix of the class which is a base for calculating accuracy measure and performance is presented below.

a	b	c	<-- classified as
23505	0	0 |	a = Undetermined
0	13225	1734 |	b = Probable
0	666	20874 |	c = Not Probable

The true undetermined cases in this confusion matrix are 23505. Those records which were predicted as true undetermined cases class by the classifier and also happened true by when tested on the test data are (true positives). The number of the records which were classified to the "Probable" class by the classifier and they are actually Probable is 13225 and 734 classified as Not Probable actually they are Probable. And also the number of instances classified to the "Not Probable" class and actually they are Not Probable on the test data are 20874 and 666 classified as probable which is actually Not Probable. 14.07%, Root Mean Square error (RMS) which indicates the mean error over each k folded testing set and 93.88% of Kappa statistic represents the mean agreement of the class predictions,

IV. Unpruned J48 Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	Undetermined
0.89	0.019	0.941	0.89	0.915	0.981	Probable
0.961	0.043	0.927	0.961	0.944	0.986	Not Probable
0.959	0.02	0.959	0.959	0.959	0.99	Weighted Avg

The confusion matrix of the class which is a base for calculating accuracy measure and performance is presented below.

a	b	c	<-- classified as
23505	0	0	a = Undetermined
0	13318	1641	b = Probable
0	834	20706	c = Not Probable

The true undetermined cases in this confusion matrix are 23505. Those records which were predicted as true undetermined cases class by the classifier and also happened true by when tested on the test data are (true positives). The number of the records

which were classified to the "Probable" class by the classifier and they are actually Probable is 13318 and 1641 classified as NotProbable actually they are Probable. And also the number of instances classified to the "NotProbable" class and actually they are NotProbable on the test data are 20706 and 834 classified as probable which is actually NotProbable.

B. J48 result indicate of type of cases

A total of two experiments (pruned and unpruned scenario) execute using J48 algorithm techniques. Pruned techniques generate a total of 506 tree size with a model accuracy of 89.2268% where as unpruned techniques generate 2722 tree size with a model accuracy of 86.4208%.

Most of the rules summarized are generate in both techniques. But the researcher focus mainly on the pruned one since it generate rule with a better accuracy and understand ability. J48 generate different interesting rules to detect malaria type of cases. The result of the experimentation summarized below.

Scenario I: Outpatient Rules Generated using J48

- **Rules that indicate uncomplicated cases**

1. if Type_of_Malaria_Visits = OutPatient and Type_of_Malaria = PV the type of cases will be UncomplicatedLabConfirmed (9402.0)
2. if Type_of_Malaria_Visits = OutPatient and Type_of_Malaria = PF the the type of cases weill be UncomplicatedLabConfirmed (9402.0)
3. if Type_of_Malaria_Visits = OutPatient and Type_of_Malaria = Notknown the type of cases will be Inpregnancy (4701.0)

Scenario II: Inpatient Rules Generated using J48

- **Rules that indicate severe anemia**

4. if Type_of_Malaria_Visits = Inpatient and Age = Under 5 the type of cases will be SevereAnemia
5. if Type_of_Malaria_Visits= Inpatient and Age=Greater 5 and Number_of_Cases >0.000004 and Rainfall <=331 and Zones
    - *All Addis Ababa zones hospital, Afar 3, S Gonder, Oromia, N Shewa, N Wello, S Wello, Waghimra, Pawe, Kemashi, Tongo, Asosa, Dire Dawa: Nuer, Gambella Town, Etang Sw, Hareri, Harerghe, Bishoftu Town, Shashemene Town, Horo Guduru, Kelem Welleganekemt Town, Jimma Town, Asela Town, Dukem Town, Sebeta Town, Mekelle and so on.*
    - Afar 1 and Number_of_Cases>6

- AWI and Rainfall <=190 and Temperature=T5, T6
- BAHIR DAR and Temperature
    - T1, T3, T4, T6, T7
    - T5 and Year (GC)=Y2004, Y2005, Y2007, Y2008
- W GOJJAM and Year (GC) =Y2004 and Occurance_of_Death=Probable

The result will be severe anemia

- **Rules indicate Inpregnancy cases**

6. if Type_of_Malaria_Visits= Inpatient and Age=Greater 5 and Number_of_Cases >0.000004 and Rainfall <=331 and Zones
    - Afar1 and Number_of_Cases<=6, Afar 2, 4 and 5, E Gojjam, Adama Town, Derashe, Alaba, Sheka, Awassa City Administration, Afder, Degehabur
    - AWI and Rainfall <=190 and Temperature =T1, T3, T4, T7
    - W GOJJAM and Year (GC) =Y2004 and Occurance_of_Death=Undeteremined:

    The type of case will be inpregnancy cases

To add more and summarize J48 rules, severe anemia cases detected when the number of cases higher and higher as well as in pregnancy case detected when the number of cases fewer and fewer. Where us occurrence of death not known (in case of outpatient cases) and probable the case will be either severe anemia or in pregnancy depends on the situation. For example:

- Metekel and Number of Cases >25: Inpregnacy or <=25 : SevereAnemia
- AGNUAK and Occurance of Death, Undetermined and Probable: SevereAnemia or NotProbable and Number of Cases <= 2.497025: SevereAnemia or > 2.497025: Inpregnancy
- GAMBELLA SW and Occurance of Death Undetermined or not probable : Inpregnancy or Probable: SevereAnemia
- MEJENGER and Number of Cases <= 2.497025: Inpregnancy or > 2.497025: SevereAnemia
- E WELLEGA and Number of Cases <= 6.464779 and Occurance of Death =Undetermined or not probable: Inpregnancy or Probable: SevereAnemia

The J48 also generated rule with the association of rain fall and number of cases. I.e. when the number of cases and rainfall less the case will be in pregnancy where us the number of rainfall less and the number of cases become higher and higher the case will be severe anemia. For example:

- JIMMA and Rainfall <= 190 and Number of Cases <= 3.497773: Inpregnancy or > 3.497773: SevereAnemia

The rule generated from J48 indicates malaria type of cases not direct association on months it varies depend on the situation related with the occurrence of deaths. For example:

- *GURAGHE and Occurrence of Deaths*
  - *Probable or outpatient cases: SevereAnemia*
  - *Occurance_of_Death = NotProbable and*
    - *Year (GC) = Y2004 and MonthNumber (GC) =*
      - *M7, M10: Inpregnancy*
      - *M8, M9, M11, M12, M1, M2, M3, M4, M5, M6: SevereAnemia*
    - *Year (GC) = Y2005 and MonthNumber (GC)*
      - *M7, M11, M12, M2, M4, M5: Inpregnancy*
      - *M8, M9, M10, M1, M2, M3, M6: SevereAnemia*

The other interesting rule generated from J48 is altitude a determinant factor for the malaria type of cases occurrence. For example:

- WOLAYTA and Altitude
  - E5, E1, E2, E3, E6: SevereAnemia (237.0)
  - E4 and Number of Cases
    - <= 6.079975: Inpregnancy
    - > 6.079975: SevereAnemia

7. Type_of_Malaria_Visits= Inpatient, Age=Greater 5 and Number_of_Cases >0.000004 and Rainfall <=331

   Regions
   - *Benishangul Gumuz, Dire Dawa, Gambella, Harari, Oromoa, SNNPR, Somali, Tigray: SevereAnemia*
8. Type_of_Malaria_Visits= Inpatient, Age=Greater 5 and Number_of_Cases >0.000004 and Rainfall >331 and Number of Cases
   - <= 4.636611: Inpregnancy or > 4.636611: SevereAnemia

II. Summary of analysis of the model (Confusion Matrix Analysis)

Table5.4. Summary of J48 type of cases models

No.	Analysis		Pruned J48	Unpruned J48
1	Tree Size	No. of Leaves	419	2305
		Size of Tree	506	2722
2	Model Accuracy	Correctly Classified Instances	89.2268%	86.4208%
		Incorrectly Classified Instances	10.7732%	13.5792%
3	Time		1.61 seconds	1.31 seconds
4	Kappa statistic		0.8344	0.7906
5	Mean absolute error		0.1947	0.0777
6	Root mean squarer error		0.1947	0.2184
7	Relative absolute error		16.573%	17.9721%
8	Root relative squared error		41.8734%	46.9818%
9	Total Number of Instances		46365	

III. Detail Accuracy by Class

The experimental evaluation of the models/rules was taken place based on the performance/accuracy of models and confusion matrix, discussion with the domain expert and based on the soundness of the rules generated.

Calculating correctly classified instances and incorrectly classified instances is based on the confusion matrix as mentioned in methodology section before. The detail of accuracy by class and confusion matrix presented below.

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	UncomplicatedLabConfirmed
0.832	0.073	0.87	0.832	0.851	0.976	SevereAnemia
0.796	0.08	0.743	0.796	0.769	0.967	Inpregnancy
0.892	0.045	0.894	0.892	0.893	0.984	Weighted Avg.

The confusion matrix of the class which is a base for calculating accuracy measures and performance is presented below.

```
a        b        c   <-- classified as
18804    0        0 |       a = UncomplicatedLabConfirmed
0        14262    2873 |    b = SevereAnemia
0        2122     8304 |    c = Inpregnancy
```

The number true uncomplicated lab confirmed cases in this confusion matrix are 18804. Those records which were predicted as true uncomplicated lab confirmed case class by the classifier and also happened true by when tested on the test data are (true positives). The number of the records which were classified to the "SevereAnemia" class by the classifier and they are actually SevereAnemia is 14262 and 2873 classified as Inpregnancy actually they are SevereAnemia. And also the number of instances classified to the "Inpregnancy" class and actually they are Inpregnancy on the test data are 8304 and 2122 classified as SevereAnemia which is actually Inpregnancy.

5.4.2. Result Generated using Multilayer Perceptron

Table 5.3 shows the level of accuracy of multilayer perceptron to indicate malaria occureence of death by attaining 86.9772 % of accuracy and accuracy to inidcate malaria probability of death occurenece.

Table 5.5. Multilayer perceptron result analysis

Parameter A	Performance
Accuracy	86.9772 %
No. Of Sigmoid Node	84
Mean absolute error	9.92%
Root mean squared error	0.26.17%
Coverage of cases (0.95 level)	96.9783 %
Time	108286.77 seconds (**more than 30 hours**)

Multilayer perceptron [**14, 42**] named as "black box" models and involve serious difficulties of theoretical interpretation that is the hidden units are essentially opaque. Even though, there are several techniques that attempt to extract rules from trained neural networks, it is unclear whether they offer any advantages over standard rule learners that induce rule sets directly from data. Thus, their utilization would only be advisable in those situations where explanation is less important than prediction [**14**].

Even if we face a diffculty to extract rule from multilayer perceptron, it identifies important variables that used of the type of cases identifications. These are age, malaria type (*PV, PF*), type of malaria visit, number of causes, rainfall, temparatue (specially when the average temprature withn in the range of 11- 20^0c and 25-30^0c), altitude (2500-3500m, 1500-2000m and 500-1000m) and months (highest in August, Januray, Febrary, April, May). Figure 5.8, 5.9 and 5.10 summarized the detail of the result of MLP.

Figure 5.8, Multilayer Perceptron Result in Occurrence of Death Classification in Month

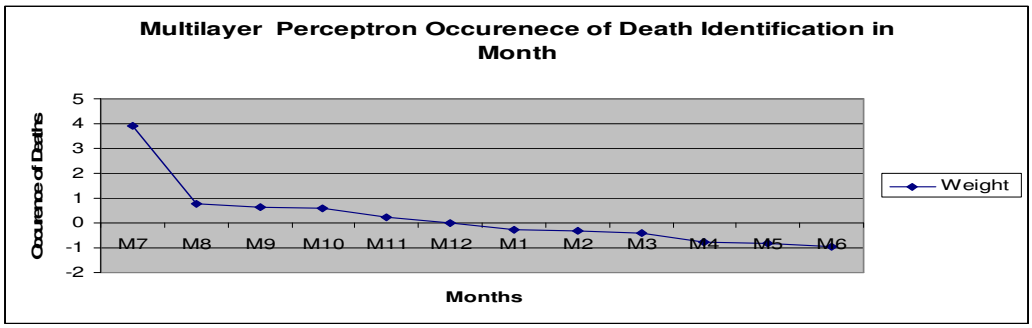


Figure 5.9, Multilayer Perceptron Result in Occurrence of Death Classification using Altitude

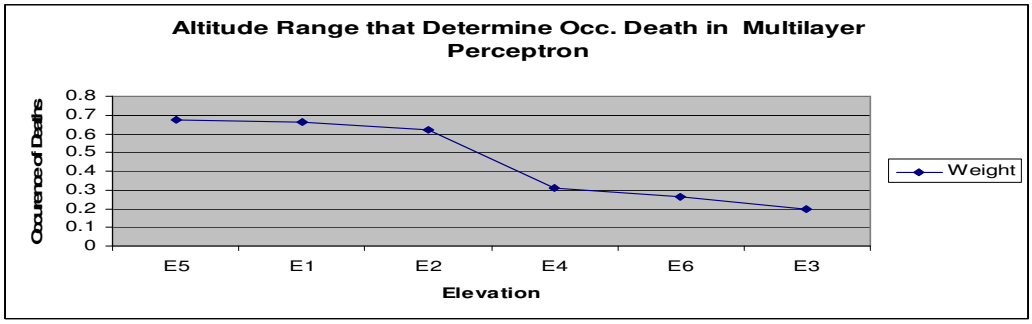


Figure 5.10, Multilayer Perceptron Occurrence of Death Classification using Temperature

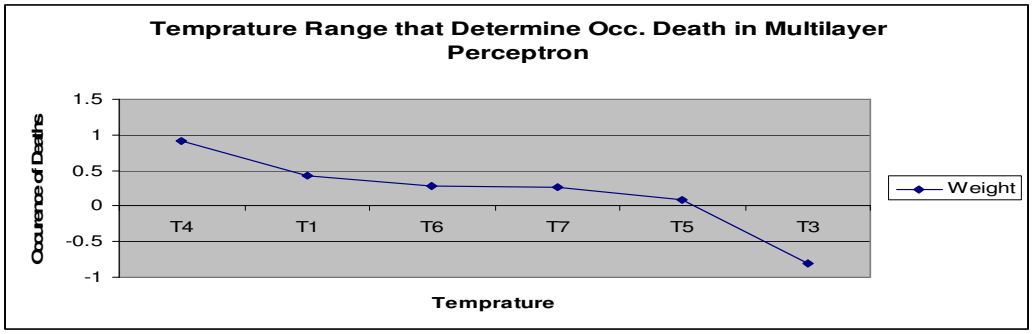


Figure 5.8, 5.9 and 5.10 indicates the probability of occurrence of deaths with related to month, altitude and temperature respectively. As shown in figure 5.8 occurrence of death is highest in M7 (July) where as probability of death almost none in month 6 (June). This may call other researcher for further investigation. In figure 5.9, occurrence of death higher with in 1000-1500m and lower in 2000-2500m. Figure 5.10 also shows occurrence of death almost none when average temperature ranges between 11-15^{0}C, where us dramatically increase in the range of 16-20^{0}c average temperature.

Table5.6. Impacts of factors to determine occurrence of deaths in multiplayer perceptron

Attribute		**Weight**	**Average Weight**
Rainfall		8.388490076	**8.388490076**
Age		0.312080754	**0.312080754**
Type_of_Malaria_Visits		-5.088496969	**-5.088496969**
Type_of_Cases	UncomplicatedLabConfirmed	0.390579598	
	SevereAnemia	0.243043173	
	Inpregnancy	0.072869046	**0.235497272**
Type_of_Malaria	PV	0.712894979	
	PF	0.274685347	
	Notknown	-0.38581972	**0.200586869**
Number_of_Cases		0.553967449	**0.553967449**

Table 5.6 clearly figures out, rainfall is high determinate factor to determine occurrence of deaths. This is may be because of its value is numeric in training dataset. Multilayer perceptron also indicates there is a high probability of occurrence of death when the type of causes severe anemia and the type of malaria is *PF*.

CHAPTER SIX
PATTERN DISCOVERY EXPERIMENTATION

Class implementing, an Apriori algorithm iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence [**50, 54**]. The algorithm has an option to mine class association rules as shown in Fig 7.1.

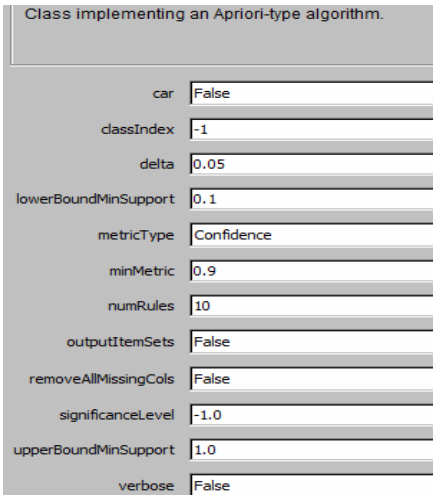


Figure 6.1 Apriori Algorithm Default Parameter Option

The detail explanation of the parameter option of Apriori algorithm listed as follows

- Car -- If enabled class association rules are mined instead of (general) association rules.
- ClassIndex -- Index of the class attribute. If set to -1, the last attribute is taken as class attribute.
- Delta -- Iteratively decrease support by this factor. Reduces support until min support is reached or required number of rules has been generated.
- LowerBoundMinSupport -- Lower bound for minimum support.
- MetricType -- Set the type of metric by which to rank rules. Confidence is the proportion of the examples covered by the premise that are also covered by the consequence (Class association rules can only be mined using confidence). Lift is confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of the importance of the association that is independent of support. Leverage is the proportion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other. The total number of examples that this represents is presented in brackets following the leverage. Conviction is another measure of departure from independence. Conviction is given by minMetric -- Minimum metric score. Consider only rules with scores higher than this value.
- NumRules -- Number of rules to find.
- OutputItemSets -- If enabled the itemsets are output as well.

- RemoveAllMissingCols -- Remove columns with all missing values.
- SignificanceLevel -- Significance level. Significance test (confidence metric only).
- UpperBoundMinSupport -- Upper bound for minimum support. Start iteratively decreasing minimum support from this value.
- Verbose -- If enabled the algorithm will be run in verbose mode.

6.1. Experimentation and Analysis of Pattern Discovery Techniques

To discover interesting pattern, we used two scenarios. On one hand using general association rules (i.e. when car value is not enabled in weka 3.7). On other hand, using class association rules (when the car value is enabled or true in weka 3.7).

A total of 120 experiments execute using Apriori algorithm techniques (60 experiments using general association rule and 60 experiments for class association rule mining) to generate the rule.

Min metric value (confidence level) the most important parameter to attain the required objective in pattern discovery. By considering this the experiment executes using 100%, 90%, 80%, 70%, 60% and 50% min metric value or confidence level. Each confidence level also experiment with a lower bound support of with a range of 10% to 100%.

For both scenarios in general association and class association rule mining the min support of the upper bound is 100% which is also the default value in weka 3.7.3. Here under we try to list the result of experimentation in detail.

6.2. General Association Rule Experiment Result

Weka 3.7 performs general association rule by using the default value of car i.e. the value of car is disabled. Here below the result of the experimentation explain in brief.

6.2.1. Experiment Scenario

Table 6.1 Scenario and Result of general association rule experiment

Confidence Level	No. of Experiments	Min Support Lower Bound	Experiment Result		
			No. Rules	No. of Cycles Performed	Min Support used to Generate the Rules
100%	5	60%-100%	No Rules Generated		
	1	50%	7	10	50%
	1	40%	7	12	40%
	3	10% - 30%	10	11	35%
90%	4	70%-100%	No Rule Generated		
	1	60%	2	8	60%
	1	50%	7	10	50%
	1	40%	7	12	40%
	3	10% - 30%	10	13	35%
80%	4	70%-100%	No Rule Generated		
	1	60%	2	8	60%
	5	10% - 50%	10	10	50%
70%	4	70%-100%	No Rule Generated		
	1	60%	2	8	60%
	5	10% - 50%	10	10	50%
60%	4	70%-100%	No Rule Generated		
	1	60%	2	8	60%
	5	10% - 50%	10	10	50%
50%	4	70%-100%	No Rule Generated		
	1	60%	2	8	60%
	5	10% - 50%	10	10	50%
Total Exp.	60				

Table 6.1 describes the number of rule generated in each confidence level. The lists of the rule generated are listed below. For example: - in table describe at 90% confidence level with min support of 60% and 20%, the techniques generate 2 and 10 rules respectively.

6.2.2. Summary of the Rules Generated from General Association Mining

General association mining strengthen the result generated using J48 and MLP. It discovers the association between occurrence of deaths, type of malaria visits, age and type of cases. Most of the rules found from the experiment support the classification model result.

6.2.3. Summary of the result

As shown in table 6.1 a total of 10 scenarios experiment (min metric or confidence level i.e. it ranges from 0.1 to 1) to generate the rule.

At 100% confidence, no rules generated when lower bound min support is 60%-100%, 90%. Based on the experiment result 10 best result generate at 100% confidence level with a minimum of support of 40% and 50% (the number of rule generated are similar except the number of cycles performed) , seven best rules and 10 best rules with a minimum support of 35% as the rule listed below.

The experiment result of Apriori techniques (Scenario II) at 90% confidence level shows that no rule generate when the minimum support (lower bound) is greater than 70% rather it should be between 35%-60%. As shown in table at 60% lower bound min support the techniques generate 2 rules, 7 rules between 40% to 50% and 10 rules at 35% min lower bound support.

Scenario III, IV, V and VI experiment performs at 80 %, 70%, 60% and 50% confidence level or min metric value. The result of all scenarios indicates no item satisfied 70% to 100% min lower bound support value. Where as from 10%-50% it generate the same result with similar cycles of performance. At 60% min support lower bound value the techniques generate two rules. How ever most of the rules are similar and as an example it looks like.

With 100% confident (expected result), for outpatient case the malaria type is uncomplicated lab confirmed and difficult to determine occurrence of death. For example with 100% confidence and 60% support:

- *If Type of Cases is Uncomplicated Lab Confirmed then Type of Malaria Visits is Outpatient*
- *If Age is Greater 5 and Occurrence of Death is Undetermined then Type of Malaria Visits is Outpatient*
- *If Age is Greater 5 and Type of Malaria Visits is Outpatient then Occurrence of Death is Undetermined*

In the case of inpatient cases how it can be the type of malaria unknown. However, the rule generates with 100% confidence level.

- *If Type of Malaria Visits is Inpatient then Type of Malaria is Not known*

6.3. Class Association Rule Experiment Result

Weka 3.7 performs class association rule by changing the default value of car i.e. when the value of car is enabled. And also similar to the previous experimentation (general association rule mining) for all scenarios the min support of the upper bound 100% used. The result of the experimentation briefly explains below.

6.3.1. Experiment Scenario

Table 6.2 Scenario and Result of class association rule experiment

Confidence Level	No. of Experiments	Min Support Lower Bound	Experiment Result		
			No. Rules	No. of Cycles Performed	Min Support used to Generate the Rules
100%	5	60%-100%	No Rules Generated		
	1	50%	3	10	50%
	1	40%	3	12	40%
	1	30%	5	14	30%
	2	10% - 20%	10	16	20%
90%	5	60%- 100%	No Rules Generated		
	1	50%	3	10	50%
	1	40%	3	12	40%
	1	30%	7	14	30%
	2	10% - 20%	10	15	25%
80%	5	60 % -100%	No Rules Generated		
	1	50%	3	10	50%
	1	40%	3	12	40%
	1	30%	7	14	30%
	2	10% - 20%	10	15	25%
70%	5	60%- 100%	No Rules Generated		
	1	50%	3	10	50%
	1	40%	3	12	40%
	1	30%	8	14	30%
	2	10% - 20%	10	15	25%
50% and 60%	5	60%- 100%	No Rules Generated		
	1	50%	3	10	50%
	1	40%	3	12	40%
	3	10% - 30%	10	14	30%
Total Exp.	60				

6.3.2. Summary of the Rules Generated from Class Association Rule

This technique also support the rule generated in general association mining. However, it discovers some different interesting rule related with temperature and malaria type. For example:

Rule 1: if Temperature is T4 and Type of Malaria Visits is Outpatient and Type of Cases is Uncomplicated Lab Confirmed then Occurrence of Death is Undetermined confidence level:(1)

- Interpretation: For outpatient cases as well as average temperature is T4 (b/n 15-20^0c), it is difficult to determine the occurrence of death with 100% confidence level.

Rule 2: if Type of Malaria is PV then Occurrence of Death is Undetermined and if Type of Malaria is PF then Occurrence of Death is Undetermined confidence level:(1)

- Interpretation: with 100% confident level it is difficult to determine occurrence of death if type of malaria either PV or PF**.** (needs further investigation whether it is unrelated or expected. However, in the case of outpatient malaria visits it's true because of the limitation of the dataset).

Rule 3: if Age is Under 5 and Type of Malaria Visits is Outpatient then Occurrence of Death is undetermined confidence level: (1)

- Interpretation: with 100% confident level, if the type of malaria visit outpatient and age is below five, it is difficult to predict occurrence of deaths.

6.3.3. Summary of the result

As shown in table 6.2 a total of 10 scenarios experiment (min metric or confidence level i.e. it ranges from 0.1 to 1) to generate the rule listed in section 7.2. The detail of the result listed in brief here below.

No rule generates when min support bound ranges between 60-70% and confidence level is 100%. In this scenario at 40%-50% min support lower bound value the Apriori techniques generate three rules and five rules. At 30%, it generates ten best rules the min support value should be in the range of 10%- 20%.

Scenario II, III and IV experiment conducts using 90%, 80% and 70% min metric value respectively and performs similar output except 8 rules generate at 70% min metric value with lower min support bound of 0.3. The result indicates, it didn't generate rule when min support lower bound above 60% where as it generate three best rule with in the range of 40%- 50%. In order to get 10 best rules using 90%

confidence value the min lower support bound should minimized in to 20%. The result indicates that the ultimate lower bound value to generate the result ranges between 25% and 50%.

When the lower bound is 0.3 the rule in scenario II, III and IV are different except the top five rules in 100% confidence level. The new rules generated at 95 confidence level using this scenario are

- if Type of Malaria Visits is Inpatient then Occurrence of Death is Not Probable and
- if Type of Malaria Visits is Inpatient and Type of Malaria is Not known then Occurrence of Death is Not Probable

Similarly when the lower bound support is below 0.2 it generates similar rule. Scenario V and VI provide similar result from the experiment and indicates the ultimate min support lower bound value should ranges between 30% and 50% to get the desired output using the confidence level of 50% and 60%. More specifically the experiment shows when the min support lower bound value above 60% no values generate, 40% and 50% it generates three rules and below 30% it generates 10 rules.

However, two different rules generates when compare with the other scenarios at confidence level 62 % and 60% respectively as follows. These are:

1. if Temperature is T4 then Occurrence of Death is Undetermined
2. if Age is Greater 5 then Occurrence of Death is Undetermined

CHAPTER SEVEN

DISCUSSION, CONCLUSION AND RECCOMMENDATION

7.1. Discussion

The researcher tried to discuss the results of J48, MLP and JRip in terms of performance of techniques (comparative evaluation of the classification rates) such as model accuracy and confusion matrix analysis. General and class association mining are also discussed in terms of support and confidence level.

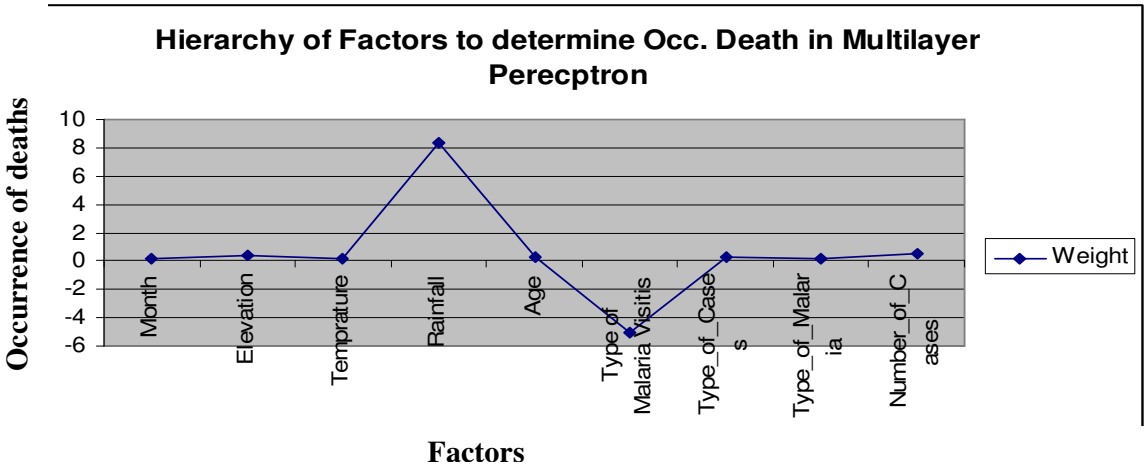
J48 discovered so many interesting rules as summarized in chapter five. On one hand, the output generates expected result in the case of outpatient malaria visits since it is difficult to extract the pattern based on the sample malaria dataset. On the other hand, most of the rules are difficult to classify based on user expectation since it requires further investigation. Factors that determine the occurrences of deaths are type of malaria visits, number of cases, age, year, month, rainfall, temperature and malaria type of cases. Based on the result of J48, as the number of cases increases there is a probability of death occurrence, especially the risk is relatively higher in less than five year olds. In most zones, the occurrence of deaths is detected between May and January since most of these zones have favorable climate conditions (such as temperature and rainfall) in these months.

J48 also discovered the most important attributes that determine the type of malaria cases. Based on the result, if the type of malaria is either PV or PF and the malaria type of visit is outpatient, the cases will definitely be uncomplicated. The most interesting rule that is impressing and need further investigation has the type of malaria unknown and the type of visits is outpatient the case will be in pregnancy. In the cases of inpatient visits, when the age is less than five years mostly the cases will be severe anemia where us when the age is greater than five it depends on the number of cases as well as the zones (with related to climate conditions). The other interesting rule generated using J48, occurrence of death mostly related with severe anemia rather than in pregnancy. The other interesting rules that call the researcher for further investigation are severe anemia cases detected when the number of cases are higher and higher as well as in pregnancy case detected when the number of cases fewer and fewer. Where us occurrence of death probability when the case is either severe anemia or in pregnancy depends on the situation.

As mentiond in [**14, 42**] clearly, the ability of multilayer perceptron to approximate non-linear functions, to filter noise in the data, etc., makes it an appropriate model to handle real problems. Nevertheless, while it is one of the most well-known and used networks, this does not imply that it is one of the most potent or that it offers the best results in different areas of application. Also computations derived from earlier input are fed back into the network, which gives them a kind of memory.

Multilayer perceptron determines the most important factors that determine the probablity of occurenece of deaths as shown in figure 7.1. Rainfall is the most significat factor that determines the occurence of death may be because of its numeric value in the training dataset. Type of malaria visist is the least determinant factor for occurence of deaths wchich completely contradicts the rule of J48 where us the other factors are almost equally important to detemine occurence of deaths.

Figure 7.1 Hierarchy of factors to determine occurrence of death in multiplayer perceptron



Apriori techniques (both general and class association mining) strengthen the result of J48 and MLP. More interestingly, it discovers the association or patterns between occurrence of deaths, type of cases and type of malaria visits.

Another point is the reduction of the performance of the system (misclassification error). This causes difference in expert and classifier prediction/judgment. Expert and classifier judgment in the knowledge discovery task is to evaluate the performance of the system in terms of how the model correctly classifies the records. In this study, the expert and classifier vary in classifying a certain records as shown in table 7.1.

Table 7.1 Misclassification error in J48

Regions		Amhara	Amhara	Amhara	Amahara
Zones		W Gojjam	W Gojjam	Bahir dar	Bahir dar
Year		2004	2004	2004	2005
Month		M9	M9	M7	M6
Altitude		E4	E4	E4	E4
Temperature		T4	T4	T4	T4
Age		Greater 5	Greater 5	Greater 5	Greater 5
Malaria visits		Inpatient	Inpatient	Inpatient	Inpatient
Malaria type		Severe anemia	Severe anemia	Severe anemia	Severe anemia
Type of cases		Not known	Not known	Not known	Not known
Number of cases		9	31	24	1
Occurrence of deaths	Actual	NotProbable	Probable	Not probable	Probable
	Predicted	Probable	Probable	Probable	Probable

Table 7.1 presents the classifier predicts the records in to a certain class as there are similar attributes that lie in the same class boundary. As shown in table 7.1 the classifier categorizes the predicted class value (occurrence of death) not probable and expert (probable) based on number of cases as well as month. The misclassification occurs in cases where the attributes are similar but one particular attribute is the most predominant in predicting the class label. However, the data represented in this research using SMOTE. As a result, the present study has mismatch between expert and classifier judgment.

In this work, the prediction method for occurrence of death and type of case prediction has been developed in a systematic way. Prediction/ classification rates of the proposed techniques show that there are no considerable differences between prediction rates of J48 and MLP. We are using neither the same type of models, nor the same implementations. Clarity in knowledge representation correctly valued in decision tree to uncover knowledge. Decision trees represent their information in a suitable form to be readable for everyone.

7.2. Conclusion

The study presents an investigation of the use of pattern discovery/prediction and data mining techniques to produce and verify malaria death/risk prediction models. A data mining methodology was proposed as the framework for use with the thesis data. The standard measurements such as mean square error, confusion matrix, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value are used to analyze the performance of the classification process.

In this paper, an empirical comparison of decision trees, rule inductions in occurrence of death or cases predictions have been presented. Our analysis extends efforts to consider issues such as knowledge description, feature selection, error distribution and computational time for training. We have found that there is not a single best alternative to cope with all these questions.

However, knowledge description of decision trees allows us to analyze interesting information such as the similarity of occurrence of deaths, cases or relevant observations. Moreover, due to the required computational time for training, decision trees could be adequate for fast prototyping in the building of interesting rules/models.

Therefore the main aim of the research was met here by uncovering the best performing scenario of data mining techniques with knowing the most determining factor/attribute for prediction of malaria deaths and cases. With this the researcher proves that applicability of data mining techniques on malaria data on the taken cases.

In general, the results from this study were encouraging. It was possible to identify the frontier determining attributes and their values for prediction and pattern discovery using data mining techniques that made good meanings to domain experts.

The decision tree with pruning found to be the top relevant technique on the dataset to get meaningful patterns from the decision tree experiments. Association rules mining was happened to be of high importance in its compactness and in bringing new attributes as a determining factors like occurrence of deaths and type of cases association/patterns which were not presented in the decision tree with pruning scenarios.

The researcher believes that a further study using data mining techniques can help to understand more about determinant factors for the death and cases of malaria. Thorough discussion with domain experts on the discovered patterns helps for getting meaningful decision support information for solving as well as improving malaria prevention and control program.

7.3. Recommendation

Further research is recommended to evaluate the effectiveness of integrating the predicting or pattern discovery model into the existing malaria control programme in terms of its impact in reducing the disease occurrence and also the cost of control interventions.

Despite the fact that the present research considers climate, elevation, location, type of malaria, type of malaria visits, number of cases and deaths attributes, the research output would help in signaling to target the malaria risk and enable focus on solving and controlling the problems. More coverage is needed in terms of economic, demographic, social and genetic factors from which the data is taken. But the researcher believes this research can be a basis for the research works done in the area of malaria dataset in future.

The next step, in this research will need the use of the knowledge based system (KBS) allowing the development of a website or multi platform executable program. So that a doctor/person can easily be informed about the malaria information in the areas with their critical factors using mobile phones or websites.

The researcher also expects to access a new set of samples collected from a new malaria dataset from the community level, in order to refine the J48 or JRip to use a different computational technique (like fuzzy logic) to create a more robust knowledge technique for malaria cases or deaths prediction and patter discovery. Furthermore, we intend to make data mining in the database used in order to find other relevant features that can be used to construct a better classifier.

Data warehouse is needed to store full historical malaria information. The data taken from different sources need integration and contains invalid values. It consumes much of the researcher time to clean and integrate.

With adequate data, GIS is very useful. Specific problem areas include accurate data on the disease and how it is reported, basic environmental and climate data such as topography, rainfall, temperature etc and demographic data on the movement of people. In the future, using the selected attributes to improve our understanding of malaria, proposed strategies used how to overcome malaria problem using GIS (GIS relevance, importance and application in controlling and understanding problem) and indicate how GIS can be used from simple mapping of malaria incidence/ prevalence all the way to sophisticated risk models. If a health district only has a digital base map and records of malaria incidence/prevalence, it should begin its use of GIS by focusing on basic mapping and not on the development of a malaria risk model.

The same work can also be done for other communicable diseases like pneumonia and diarrhoeal diseases, trypanosomiasis, TV, onchocerciasis, HIV/AIDS and so .on

Reference

1. WHO Report (2009), http://www.who.int/topics/malaria/WHOmalaria.htm, [Access date August 07, 2010].
2. Perlino,C.(2006), "The public health work force: Left unchecked, will we be protected?", American Public Health Association.
3. USAID report (2004), http://www.usaid.gov/our_work/global_health/id/malaria/; [Accessed September 07, 2010].
4. UNICEF Report (2010), http://www.unicef.org/health/index_malaria.html, [Accessed August 07, 2010].
5. Federal Democratic Republic of Ethiopia Ministry of Health (2008), "Ethiopia National Malaria Indicator Survey 2007 Technical Summary".
6. Ethiopian Federal Democratic Republic Ministry of Health (2010), "Malaria Operational Plan Ethiopia", Endorsed by the U.S. Global Malaria Coordinator.
7. Central Statistical Agency (2005), "Ethiopia Demographic and Health Survey", Addis Ababa, Ethiopia ORC Macro Calverton, Maryland, USA
8. Aynalem Adugna (2007), "Malaria in Ethiopia", http://www.EthioDemographyAndHealth.Org, [Accessed September 07, 2010].
9. Ethiopian Federal Democratic Republic Ministry of Health (2006), "National Malaria Control Program, Malaria Prevention and control in Ethiopia", Addis Ababa.
10. Glenn J. Myatt (2007), "A Practical Guide to Exploratory Data Analysis and Data Mining", John Wiley & Sons.
11. Han Jiawei and Kamber Micheline (2006), "Data Mining: concepts and Techniques", Morgan kufman Publishers.
12. Thearling et.al (2003), "An Overview of Data Mining Techniques", [Accessed September 04, 2010]
13. Ethiopian Federal Democratic Republic of Ethiopia Ministry of Health (2009/2010), "Health and Health Related Indicators", Policy Plan and Finance General Directorate, Page 30-31, 43 and 44
14. Witten, I. H. and Frank, E. (2005), "Data mining: practical machine learning tools and techniques", Morgan Kaufmann series in data management systems, Morgan Kaufman Publishers.

15. Teklu Urgessa (2010), "Application of Data Mining Techniques on Antiretroviral Therapy (Art) Data: The Case of Adama and Asella Hospitals", MSC Thesis, Addis Ababa University, Ethiopia.
16. Shagaw Anagew (2002), "Application of Data Mining Technology to Predict Child Mortality Pattern: The Case of Butajira Rural Health Project", MSC Thesis, Addis Ababa University, Ethiopia.
17. Abraham Tesso (2005), "Application of Data Mining Technology to Identify Determinant Risk Factors of HIV Infection to Find Their Association Rules: The Case of Center for Disease Controls and Prevention (CDC)", MSC Thesis, Addis Ababa University, Ethiopia.
18. Mirjam Schunk, Wondimagegn P Kumma, Isabel Barreto Miranda, Maha E Osman, Susanne Roewer, Abraham Alano, Thomas Löscher, Ulrich Bienzle and Frank P Mockenhaupt (2006), "High prevalence of drug-resistance mutations in Plasmodium falciparum and Plasmodium vivax in southern Ethiopia", BioMed Central Ltd.
19. Kaliyaperumal Karunamoorthiab and Mammo Bekelea (2009), "Prevalence of malaria from peripheral blood smears examination: A 1-year retrospective study from the Serbo Health Center, Kersa Woreda, Ethiopia", Journal of Infection and public health.
20. Mehmed Kantardzic (2003), "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons.
21. David Hand, Heikki Mannila, and Padhraic Smyth (2001), "Principles of Data Mining", MIT Press, Cambridge, MA.
22. Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi (1998), "Discovering Data Mining: From Concept to Implementation", Prentice Hall, Upper Saddle River, NJ.
23. S. Sumathi and S.N. Sivanandam(2006) , "Introduction to Data Mining and its Applications", Springer.
24. Daniel T. Larose (2005), "Discovering Knowledge in Data: An Introduction to Data Mining," John Wiley & Sons.
25. Peter Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinart, Colin Shearer, and Rudiger Wirth (2000), "CRISP–DM Step-by-Step Data Mining Guide", http://www.crisp-dm.org/ [accessed October 5, 2010].

26. Daud M.N.R., and Corne, D.W., (2007), "Human Readable Rule Induction in Medical Data Mining: A Survey of Existing Algorithms", WSEAS European Computing Conference, 2007, Athens, Greece.
27. Cornelia Gyorödi, Robert Gyo rödi and Stefan Holban (2003), "A Comparative Study of Association Rules Mining Algorithms", University of Oradea and Politehnica University of Timisoara, Romania, http://www.cs.utt.ro/~stefan, [accessed February 02, 2011].
28. A.K. Jain, M.N. Murty and P.J. Flynn (2000), "Data Clustering: A Review", Michigan State University ,Indian Institute of Science and The Ohio State University, ACM Computing Surveys, Vol. 31,Page 3
29. Mary K. Obenshain (2004), "Application of Data Mining Techniques to Healthcare Data, Data Quality Research Institute", UNC
30. Shoji Hirano and Shusaku Tsumoto (2007), "Temporal Data Mining in Hospital Information Systems: Analysis of Clinical Courses of Chronic Hepatitis", TSI USA, Vol. 1, Page 11-19.
31. Austeclino Magalhaes Barros Junior and Angelo Amancio Duarte (2010), "Artificial Neural Networks and Bayesian Networks as Supporting Tools for Diagnosis of Asymptomatic Malaria", IEEE.
32. Kinley Wangdi , Pratap Singhasivanon , Tassanee Silawan, Saranath Lawpoolsri , Nicholas J White and Jaranit Kaewkungwal (2010), "Development of temporal modeling for forecasting and prediction of malaria infections using time-series and ARIMAX analyzes: A case study in endemic districts of Bhutan", Malaria Journal.
33. Arantxa Roca –Feltrer, , Ilona Carnei ro, Lucy Smith, Joanna RM Armstrong Schellenb erg, Brian Greenwood and David Schell enberg (2010), "The age patterns of severe malaria syndromes in sub-Saharan Africa across a range of transmission intensities and seasonality settings"; Malaria Journal.
34. Yazoumé Yé, Valérie R Louis, Séraphin Simboro and Rainer Sauerborn (2007),"Effect of meteorological factors on clinical malaria risk among children: an assessment using village-based meteorological stations and community-based parasitological survey", BioMed Central Ltd.
35. Hailay D TeklehaimanotMarc, Lipsitch, Awash Teklehaimanot and Joel Schwartz (2004); "Weather-based prediction of Plasmodium falciparum malaria in

epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms", Malaria Journal.

36. Robert D. Newman, Afework Hailemariam, Daddi Jimma, Abera Degifie, Daniel Kebede, Aafje E. C. Rietveld, Bernard L. Nahlen, John W. Barnwell, Richard W. Steketee and Monica E. Parise (2003), "Burden of Malaria during Pregnancy in Areas of Stable and Unstable Transmission in Ethiopia during a Non epidemic Year", Journal of Infectious Diseases, University of Chicago Press.
37. Natacha Protopopoff, Wim Van Bortel, Niko Speybroeck, Jean-Pierre Van Geertruyden, Dismas Baza, Umberto D'Alessandro and Marc Coosemans (2009), "Ranking Malaria Risk Factors to Guide Malaria Control Efforts in African Highlands", United States of America.
38. Asnakew K Yeshiwondim, Sucharita Gopal, Afework T Hailemariam, Dereje O Dengela and Hrishikesh P Patel (2009)**, "** Spatial analysis of malaria incidence at the village level in areas with unstable transmission in Ethiopia", BioMed Central Ltd.
39. Arantxa Roca-Feltrer, Joanna RM Armstrong Schellenberg, Lucy Smith and Ilona Carneiro (2009), "A simple method for defining malaria seasonality", Malaria Journal.
40. Ana Azevedo and Manuel Filipe Santos (2008), "KDD, SEMMA AND CRISP-DM: A Parallel Overview", IADIS.
41. Lukasz A. Kurgan and Petr Musilek (2006), A survey of Knowledge Discovery and Data Mining process models, The Knowledge Engineering Review", Cambridge University Press, Vol. 21, Page 1–24.
42. Zuleyka Díaz, María Jesús Segovia, José Fernández and Eva María del Pozo (2005), "Machine Learning and Statistical Techniques. An Application to the Prediction of Insolvency in Spanish Non-life Insurance Companies", The International Journal of Digital Accounting Research Vol. 5, Page 9
43. Krzysztof J. Cios and G. William Moore (2002), "Uniqueness of Medical Data Mining", Artificial Intelligence in Medicine journal.
44. Malaria Consortium (2006), "Malaria: a Handbook for Health Professionals", Macmillan.
45. David G. Kleinbaum, Kevin M. Sullivan and Nancy D. Barker (2007), "A Pocket Guide to Epidemiology", Springer.
46. Malaria consortium (2006), "District handbook for malaria control".

47. WHO Regional Office for Africa (2001), "Integrated Disease Surveillance in the African Region: A regional Strategy for Communicable Diseases".
48. Sarah Boslaugh (2007), "Secondary Data Sources for Public Health: A Practical Guide", Cambridge University Press.
49. J. Ross Quinlan (1994), "C4.5: Programs for Machine Learning", Morgan Kaufmann.
50. Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg (2007), "Top 10 algorithms in data mining", Springer.
51. Sam Drazin and Matt Montag (2000), "Decision Tree Analysis Using Weka", Machine Learning- Project II, University of Miami
52. University of Waikato (2008), "Weka Manual Version 3.6".
53. Cornelia Gyorödi, Robert Gyo rödi and Stefan Holban (2004) , "A Comparative Study of Association Rules Mining Algorithms", University of Oradea, Str. Armatei Romane, Department of Computer Science and Politehnica University of Timisoara, Department of Computer and Software Engineering
54. Bing Liu, Wynne Hsu, Yiming Ma (1998), "Integrating Classification and Association Rule Mining", Fourth International Conference on Knowledge Discovery and Data Mining, Page 80-86.
55. Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W (2002), "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research.
56. Juanjuan Wang, Mantao Xu, Hui Wang and Jiwu Zhang (2006), "Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding", IEEE.
57. Nitesh V. Chawla, Nathalie Japkowicz and Aleksander Kolcz (2004), "Special Issue on Learning from Imbalanced Data Sets", Sigkdd Explorations, Vol.6, Page 2.
58. Gustavo E. A. P. A. Batista, Ronaldo C. Prati and Maria Carolina Monard (2004), "A Study of the Behaviour of Several Methods for Balancing Machine Learning Training Data", Sigkdd Explorations, Vol. 6, Page 21.

Annex 1. J48 Generated rules using Weka

1. Type_of_Malaria_Visits = OutPatient: Undetermined (23505.0)

2. Type_of_Malaria_Visits = Inpatient and Number_of_Cases <= 0.00001: NotProbable (10933.49/48.0)

3. Type_of_Malaria_Visits = Inpatient and Number_of_Cases <= 12.01961 and | | Number_of_Cases > 0.00001 and zone

- Lideta, N Gonder, S Gonder, Oromia, N Shewa, N Wello, S Wello, Wag Himra, Pawe, Metekel, Kemashi, Tongo, Asosa, Dire Dawa:, Agnuak, Gambella Sw, Nuer, Mejenger, Gambella Town, Etang Sw, Hareri, E Wellega, E Harerghe, W Wellega, Jimma, Illubabor, E Shewa,Akaki, Gulele, Yeka, Bole, Addis Ketema, Kolfe-Keranio, Arada, Nefas Silk Lafto, Arada, Zewditu Hospital, Yekatit Hospital, Alert Hospital, Menilik Hospital, Ras Desta Hospital, Tikur Anbessa Hospital, Afar 1, Afar 2, Afar 3, Afar 4, Afar 5, Afar 1, Awi: Borena, Bishoftu Town, Shashemene Town, W Arsi, Kelem Wellega, Nekemt Town, Jimma Town, Asela Town, Burayu, Adama Town, Dukem Town, Sebeta Town, Gedeo,Gamo Gofa: Derashe: Alaba: Kembata/Tembabo: Konso: Yem: Silti: Hadiya:S Omo: Bench Maji: Wolayta: Sidama: Sheka: Konta: Burji: Awassa City Administration: Jijiga: Afder: Gode: Warder: Degehabur: Korahe: Fik: Shinile: Sw Tigray: Nw Tigray: C Tigray: E Tigray: S Tigray: Sw Shewa, W Shewa, W Harerghe, Guji, Arsi: **NotProbable**
- Cherkos and Year (GC)
  - Y2004 and Rainfall
    - ✓ <=5: NotProbable
    - ✓ >5: Probable
  - Y2005, Y2006, Y2007, Y2008: NotProbable
- BAHIR DAR and Number of Cases
  - > 1.117114 and <= 1.18504: Probable (10.0/3.0)
  - <= 1.195529 and > 1.18504: Probable (8.0/0.0)
  - <= 4.803137 and > 1.557748 and
    - ✓ Age = Greater 5: Probable (649.0/260.0)
- W GOJJAM and MonthNumber (GC)
  - M11 and Year (GC) = 2004 and Number of Cases
    - ✓ <= 6.426978: Probable (245.02/10.02)
    - ✓ > 6.426978 and <= 11.426374: Probable (192.02/76.02)
  - M11 and Year GC(2005) and Number of Cases

- ✓ <= 4.39663: Probable (1644.14/3.14)
- ✓ > 4.39663 and <= 9.099446: Probable (640.05/25.05)

- The Rest of year no probability of deaths

- BALE and Temperature
  - T5 and Number of Cases
    - ✓ <= 3: Probable (4.0/0.0)
- HORO GUDURU
  - Age = Under 5: NotProbable (3.0)
  - Age = Greater 5: Probable (3.0/1.0)
- GURAGHE and Temperature
  - T4 and Age = Under 5 and Rainfall
    - ✓ > 185 and Number_of_Cases > 3: Probable (4.04/0.04)
  - Age = Greater 5: NotProbable (1347.0/3.0)
  - T5 and Type of Cases
    - ✓ SevereAnemia and Age = Greater 5: Probable (21.0/6.0)
  - No Death Probablities in other temperature ranges
- KEFFA and Number of Cases
  - > 1.495456: Probable (3.0/1.0)
- AMARO and MonthNumber (GC)
  - M9 and M10: Probable (2.0/0.0)
- DAWRO and Age
  - Under 5: Probable (6.41/1.41)
- BASKETO and Number of Cases
  - > 3.496872: Probable (11.0/5.0)
- LIBEN and Number of Cases
  - > 5.49636: Probable (3.06/1.06)
- MEKELLE and Age
  - Under 5 and Year
    - ✓ Y2004: NotProbable (6.0)

4. Type_of_Malaria_Visits = Inpatient and Number_of_Cases > 12.01961 and Regions

- AFAR and Age = Under 5: NotProbable (6.88)
- AMHARA and Year (GC) = Y2004 and Number of Cases
  - <= 12.484261 and Rainfall <= 15.9425 and Rainfall > 15.001088: Probable (4.8)

- > 19.601178 and <= 21.30256 and Rainfall <= 15.9425 and > 15.001088 : Probable (17.0/1.4)
- Zones = N GONDER: Probable (4.0/1.0)
- > 19.164967: Probable (7482.0/8.0)

- DIRE DAWA: Probable (0.0)
- SNNPR and Number of Cases
  - > 17.957834 and Type_of_Cases = UncomplicatedLabConfirmed: Probable (0.0)
  - > 17.957834 and Type_of_Cases = SevereAnemia and Age = Under 5 and Temprature
    - ✓ T5 and Number_of_Cases > 38.499341: Probable (10.0/3.0)
  - > 17.957834 and Type_of_Cases = SevereAnemia and Age = Greater 5: Probable (55.0/22.0)
- SOMALI: Probable (3.18/1.18)
- TIGRAY and Age
  - Age = Greater 5 and Year
    - ✓ Y2004: Probable (2.0/1.0)
    - ✓ The reset not probable

Annex 2. JRip rule generated using Weka

Rule 1

- (MonthNumber (GC) = 11) and (Zones = W GOJJAM) and (Year (GC) = Y2005) => Occurance_of_Death=Probable (8619.0/128.0)

Rule 2

- (MonthNumber (GC) = M11) and (Number_of_Cases >= 12.095628) and (Type_of_Cases = SevereAnemia) and (Zones = BAHIR DAR) => Occurance_of_Death=Probable (2912.0/25.0)

Rule 3

- (MonthNumber (GC) = 11) and (Zones = W GOJJAM) and (Number_of_Cases >= 21.358824) and (Number_of_Cases >= 29.173761) and (Type_of_Malaria_Visits = Inpatient) => Occurance_of_Death=Probable (663.0/2.0)

Rule 4

- (MonthNumber (GC) = 11) and (Zones = W GOJJAM) and (Number_of_Cases <= 6.796859) => Occurance_of_Death=Probable (260.0/17.0)

Rule 5

- (MonthNumber (GC) = M11) and (Regions = AMHARA) and (Zones = W GOJJAM) and (Number_of_Cases >= 19.638148) and (Number_of_Cases <= 28.195373) => Occurance_of_Death=Probable (308.0/43.0)

Rule 6

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.803137) and (Number_of_Cases >= 0.070563) and (Number_of_Cases >= 1.559582) and (Year (GC) = Y2005) and (Number_of_Cases <= 3.656529) and (Number_of_Cases >= 3.248916) => Occurance_of_Death=Probable (72.0/14.0)

Rule 7

- (MonthNumber (GC) = MM1M1) and (Regions = AMHARA) and (Zones = W GOJJAM) and (Number_of_Cases >= 17.383078) and (Number_of_Cases <= 18.428765) => Occurance_of_Death=Probable (28.0/1.0)

Rule 8

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.803137) and (Number_of_Cases >= 0.056115) and (Number_of_Cases >= 1.522661) and (Number_of_Cases >= 4.0347) => Occurance_of_Death=Probable (127.0/45.0)

Rule 9

- (MonthNumber (GC) = MM1M1) and (Regions = AMHARA) and (Zones = W GOJJAM) and (Number_of_Cases <= 11.413941) => Occurance_of_Death=Probable (183.0/75.0)

Rule 10

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.019931) and (Number_of_Cases >= 0.043085) and (Number_of_Cases >= 1.522661) and (Number_of_Cases <= 1.781393) => Occurance_of_Death=Probable (53.0/12.0)

Rule 11

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.803137) and (Number_of_Cases >= 0.055064) and (Number_of_Cases >= 2.036494) and (Number_of_Cases <= 2.864673) and (Number_of_Cases <= 2.119483) => Occurance_of_Death=Probable (23.0/1.0)

Rule 12

- (MonthNumber (GC) = MM1M1) and (Regions = AMHARA) and (Number_of_Cases <= 4.803137) and (Number_of_Cases >= 0.055064) and (Number_of_Cases >= 2.208702) and (Number_of_Cases <= 2.969485) and (Number_of_Cases >= 2.627885) => Occurance_of_Death=Probable (60.0/16.0)

Rule 13

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.019931) and (Number_of_Cases >= 0.055064) and (Number_of_Cases <= 0.986747) and (Number_of_Cases >= 0.268382) and (Number_of_Cases <= 0.417283) and (Number_of_Cases >= 0.380814) => Occurance_of_Death=Probable (32.0/7.0)

Rule 14

- (MonthNumber (GC) = MM1M1) and (Regions = AMHARA) and (Number_of_Cases >= 7.98208) and (Type_of_Cases = SevereAnemia) and (Number_of_Cases >= 11.654699) and (Number_of_Cases <= 12.259415) and (Number_of_Cases >= 11.997793) => Occurance_of_Death=Probable (27.0/3.0)

Rule 15

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.019931) and (Number_of_Cases >= 0.056115) and (Number_of_Cases >= 0.488562) and (Number_of_Cases <= 0.626681) and (Number_of_Cases <= 0.533688) and (Number_of_Cases >= 0.519549) => Occurance_of_Death=Probable (21.0/2.0)

Rule 16

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.019931) and (Number_of_Cases >= 0.022803) and (Number_of_Cases >= 0.069397) and (Number_of_Cases <= 0.114539) and (Number_of_Cases >= 0.088452) => Occurance_of_Death=Probable (88.0/33.0)

Rule 17

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.019931) and (Number_of_Cases >= 0.033897) and (Number_of_Cases >= 2.145702) and (Number_of_Cases <= 2.505152) and (Number_of_Cases >= 2.208702) => Occurance_of_Death=Probable (62.0/20.0)

Rule 18

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.019931) and (Number_of_Cases >= 0.022803) and (Type_of_Cases = SevereAnemia) and (Number_of_Cases >= 0.5009) and (Number_of_Cases <= 1.033505) and (Number_of_Cases >= 0.786183) => Occurance_of_Death=Probable (96.0/41.0)

Rule 19

- (Zones = BAHIR DAR) and (Number_of_Cases <= 4.019931) and (Number_of_Cases >= 0.022803) and (Number_of_Cases >= 0.268382) and (Number_of_Cases <= 0.626681) and (Number_of_Cases <= 0.309922) and (Number_of_Cases >= 0.298453) => Occurance_of_Death=Probable (13.0/0.0)

Annex 3: Generated rule using patteron discovery in weka

1 Occurance_of_Death = Undetermined 23505 ==> Type_of_Malaria_Visits = OutPatient 23505 conf:(1)

2 Type_of_Malaria_Visits=OutPatient 23505 ==> Occurance_of_Death = Undetermined 23505 conf:(1)

3 Type_of_Cases=UncomplicatedLabConfirmed 18804 ==> Type_of_Malaria_Visits = OutPatient 18804 conf:(1)

4 Type_of_Cases=UncomplicatedLabConfirmed 18804 ==> Occurance_of_Death = Undetermined 18804 conf:(1)

5 Type_of_Cases=UncomplicatedLabConfirmed Occurance_of_Death = Undetermined 18804 ==> Type_of_Malaria_Visits=OutPatient 18804 conf:(1)

6 Type_of_Malaria_Visits=OutPatient Type_of_Cases = UncomplicatedLabConfirmed 18804 ==> Occurance_of_Death=Undetermined 18804 conf:(1)

7 Type_of_Cases=UncomplicatedLabConfirmed 18804 ==> Type_of_Malaria_Visits =OutPatient Occurance_of_Death=Undetermined 18804 conf:(1)

8 Type_of_Malaria_Visits=Inpatient 14103 ==> Type_of_Malaria=Notknown 14103 conf:(1)

9 Age=Greater 5 Occurance_of_Death=Undetermined 14103 ==> Type_of_Malaria_Visits =OutPatient 14103 conf:(1)

10 Age=Greater 5 Type_of_Malaria_Visits=OutPatient 14103 ==> Occurance_of_Death =Undetermined 14103 conf:(1)

11 Temprature=T4 20327 ==> Occurance_of_Death=Undetermined 12699 conf:(0.62)

12 Age=Greater 5 23505 ==> Occurance_of_Death=Undetermined 14103 conf:(0.6)